

(19)日本国特許庁(JP) (12)公開特許公報(A)

(11)特許出願公開番号

特開2003-122509

(P2003-122509A)

(43)公開日 平成15年4月25日(2003.4.25)

(51)Int.Cl.	G 0 6 F	3/06	12/00	5 3 1	5 3 3	F I	G 0 6 F	3/06	12/00	5 3 1 D	5 B 0 8 2	6 3 3 J	チーゴト(参考)

審査請求 未請求 請求項の数34 OL (全 34 頁)

(21)出願番号	特開2002-19971(P2002-19971)	(71)出願人	00005108
(22)出願日	平成14年1月29日(2002.1.29)	株式会社日立製作所	
(31)優先権主張番号	特開2001-240072(P2001-240072)	東京千代田区神田駿河台四丁目6番地	
(32)優先日	平成13年8月8日(2001.8.8)	中野 俊夫	
(33)優先権主張国	日本(JP)	神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内	
		中村 勝彦	
		神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内	
		100071283	
		弁理士 一色 健輔	

最終頁に続く

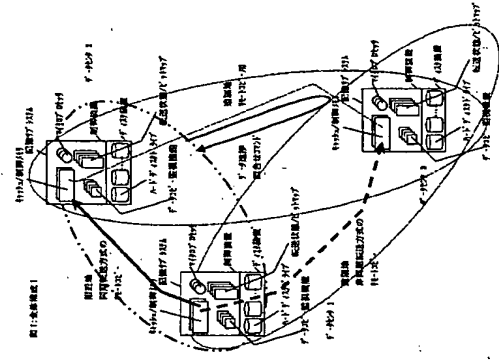
(54)【発明の名称】 リモートコピー制御方法、これを用いた記憶サブシステム、及び、これらを用いた広域データ...

トレージシステム

(57)【要約】 (修正有)

【課題】 3以上のデータセンタ間で、常時、データ更新の順序性を保持する。

【解決手段】 付近地に存在する2つのデータセンタ間では同期転送によるコピー機能を用いた接続構成とする。これらのうち1つのデータセンタと、遠隔地に存在する第3のデータセンタとの間は、非同期リモートコピー機能で連結し、付近地に存在する記憶サブシステムがホストから受領したデータの順序性を常時、保証しつつ第3のデータセンタが、そのデータを保持する。更に、正常運用の際には直接にデータ転送を行わない2つのデータセンタに設置された記憶サブシステムの間で、データ転送・受領・更新の進捗状態を把握する機能を、各記憶サブシステムに設ける。



第2の記憶サブシステム、及び、第3のデータセンタに設置された第3の記憶サブシステムを有する広域データストレージシステムにおける、3つ以上のデータセンタの間のリモートコピー制御方法であって、

する広域データストレージシステムにおけるリモートコピー制御方法において、
第1の記憶サブシステムが、前記ホストからの指示により第1の記憶装置に対してデータの書き込みを行うステップと、

る第1のステップと、
第1の記憶サブシステムが、前記ホストからのデータ
を、第2の記憶サブシステムへ、同期転送する第2のステ
ップと、

第1の記憶サブシステムが、第1の記憶資源における前記データ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号との対応づけとを記憶するステップと、

第2の記憶サブシステムが、同期転送された前記ホストからのデータを、第3の記憶サブシステムへ、非同期転送する第3のステップと、

10 第1の記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記対応づけたを第2の記憶サブシステムに送信するステップと、

第1の記憶サブシステムから第3の記憶サブシステムへ、前記ホストからのデータが第3の記憶サブシステムへ到着したか否かを問合せる第4のステップとを有するリモートコピー制御方法。

第2の記憶サブシステムが、前記データと前記対応づけ
とを受信して前記データを第2の記憶資源に記憶し、前
記データと前記対応づけとを第3の記憶サブシステムに
送信するステップと、

【請求項13】 請求項12記載のリモートコピー制御方法において、更に、第2のデータセンタが機能停止する第5のステップと、

第3の記憶サブシステムが、前記データと前記対応づけ
とを受信して前記データを第3の記憶資源に記憶すると
ともに前記対応づけにおける前記シーケンス番号を第1

第1の記憶サブシステムから第3の記憶サブシステムへ、第1の記憶サブシステムが保持するデータの一部を転送する第6のステップとを有するリモートコピー制御方法。

20 の配座サブシステムに送信するステップと、
第1の配座サブシステムが、前記シーケンス番号を受信
して、前記シーケンス番号と記憶している前記対応づけ
とに基づいて、第3の記憶資源に未反映となっているデ
ータを記憶するステップと、

を備え、前記記憶資源に記憶されているデータが転送される複数の転送先が接続する記憶サブシステムにおけるモータコピ制御方法において、

25 を有することを特徴とするリモートコピー制御方法。

【請求項16】 前記第2の記憶サブシステムが障害等により使用できなくなった場合に、

記憶サブシステムが、前記憶資源にデータを書き込む
 ステップと、

前記第 1 の記憶サブシステムが、前記受信した前記システム番号と前記対応づけとに基づいて把握した前記第 3 の記憶資源において未反映となっている差分のデータとその書き込み位置情報とを前記第 3 の記憶サブシステムに送付するステップと、

の対応づけとを記憶するステップと、

前記第3の記憶サブシステムが、前記差分データと前記

記憶サブシステムが、前記データ書き込みにより書き込んだデータと前記対応づけとを前記転送先に送信するステップと。

書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて前記第3の記憶資源に記憶して前記第1の記憶資源と前記第3の記憶資源の内容を同期させるステップと、

シーケンス番号を受信するステップと、
記憶サブシステムが、記憶している前記対応づけと前記
転送先から受信した前記シーケンス番号とに基づいて、
前記転送先において未反映となっている書き込みデータ
を把握するステップと

第1の記憶サブシステムが、前記ホストからの指示により前記第1の記憶資源に対してデータ書き込みが行われた場合に、前記データ書き込みにより書き込まれたデータと前記第1の記憶資源における前記データ書き込みが行われた位置とを特定する書き込み位置情報とを前記第3の記憶サブシステムに送信する。

を有することを特徴とするリモートコピー制御方法。
 【請求項15】 第1の記憶資源に対するデータ書き込み手段を備え第1のサイトに設置された第1の記憶サブシステムと、第1の記憶サブシステムに接続するホストシステムと、第2の記憶資源に対するデータ書き込み手段を備え、第2のサイトに設置された第2の記憶サブシステムと、第3の記憶資源に対するデータ書き込み手段を備え第3のサイトに設置された第3の記憶サブシステムと、を有する。

の記憶サブシステムに送信するステップと、
前記第3の記憶サブシステムが、前記データと前記書き
込み位置情報とを受信して前記データを前記第3の記憶
装置の前記書き込み位置情報で特定される位置に記憶す
るステップと、
を有することを特徴とする請求項15に記載のリモート
コピー制御方法。

第3の記憶サブシステムから送られてくるシーケンス番号を受信して、このシーケンス番号と、自身が記憶している前記対応づけとに基づいて、第3の記憶資源に未反映となつていないデータを把握する手段と、
 55 備えることを特徴とする記憶サブシステム。
 【請求項34】 ホストが接続された記憶サブシステムAとこれに通信手段を介して接続する複数の他の記憶サブシステムBと、を有する広域データストレージシステムにおける請求項24に記載のリモートコピ制御方法に使用する前記記憶サブシステムBとして機能する記憶サブシステムであって、
 10 前記ホスト、記憶サブシステムA、前記通信手段のうち少なくともいずれか一つに障害が発生した場合に障害を検出した場合に記憶サブシステムBの中から記憶サブシステムAの処理を代行させる記憶サブシステムBを導出する手段と、
 15 導出した記憶サブシステムbとこれ以外の記憶サブシステムBが記憶しているデータの内容を一致させる手段と、
 20 記憶サブシステムbに接続する副ホストにより前記ホストの運用を引き継ぐ手段と、
 25 【発明の詳細な説明】
 【0001】
 【発明の属する技術分野】 本発明は、災害による外部記憶装置の障害が生じた後に、速やかに、その障害から復旧可能な広域データストレージシステムに依り、特に、外部記憶装置が相互に100kmから数百km隔てて設置され、相補的な動作を行う3つ以上の外部記憶装置からなる広域データストレージシステムに関する。
 【0002】
 【従来の技術】 本件の特許人による特開平11-338647号公報には、システムとデータの多重化を同期又は非同期に行うことが開示されている。また、本件の特許人による特開2000-305856号公報には、非同期で遠隔地にデータのコピーを行う技術が開示されている。

【0003】このように、本件の特許人は、大型計算機システム、サーバ、ネットワーク上のパーソナルコンピュータ、その他の上位計算機システム（以下、ホストという。）から、データの順序を特定する特別な制御情報を受領することなく、そのデータを受け取った外部記憶装置（以下、記憶サブシステムという。）が、そのデータを、遠隔地に設置された第2の記憶サブシステムに対し、そのデータの順序性を常時保証しながら、非同期転送により第2の記憶サブシステムへ連続して断続なく書き込むという、非同期リモートコピの技術を所有している。
 【0004】また、同期転送の技術を用いてコピーを行うと、同期転送の技術を用いてコピーを行

うときは、ホストとこれに接続された記憶サブシステムとの間のデータ更新処理と、この記憶サブシステムと付近地又は遠隔地に設置された記憶サブシステムとの間のコピ制御が運動するため、巨额的にみて常に2つの記憶サブシステム間でデータが一致しており、その書き込み順序性も同時に保証されている。尚、適当なデータ転送経路を選択すれば、2つの記憶サブシステムの距離が100kmを超える場合であっても、同期転送によるコピーが可能である。

【0005】昨今、データを安全に格納し保持することが重要であるという認識が高まっており、データストレージの市場では、ディザスタリカバリシステムを要請する声が多く聞かれる。従来のように、データ格納地点を2つ設け、かかる2地点間を同期転送又は非同期転送で結ぶことは実現されている。しかし市場は、第3、第4のデータ格納地点（以下、データセンターという。）を要求し、これらの間の完全な又は完全に近いディザスタリカバリシステムの実装を望んでいる。
 【0006】その理由は、3地点以上のデータセンターを設置しておけば、これらのうち1箇所が災害に見舞われても、引き続き発生する災害のリスクを軽減するため、に残る複数のセンター間でデータ冗長化の回復・維持が図れるであろうという期待にある。

【0007】従来の技術では、3以上のデータセンターを構築した場合に、ホストから受領する1/0を唯一の記憶サブシステムの論理ボリュームで受領し、これを複数のデータセンターへリモートコピ技術を用いて転送する際の配慮が十分で無かった。例えば、一つのデータセンターが災害によりダウンした場合に、残る2以上のデータセンター間で、データの順序性を保証した論理ボリュームを構築できるか、更新状態を引き継ぎデータの不整合を無くすることができるか、附近地又は遠隔地に對するコピーを可能とするシステムの再構築ができるかといった問題に関し、配慮が足りなかった。

【0008】
 【発明が解決しようとする課題】 災害はいつ発生するか不明なため、3以上のデータセンター間で、常時、データ更新の順序性を保持しなければならない。
 【0009】このため、ホストに特殊な機能を用意せず、複数のリモートコピ構成を連結し、同一論理ボリュームが受領したデータを、遠隔地又は附近地の別の記憶サブシステムへ配信し、かつ、如何なる時点で災害が発生しても、ホストからのデータの更新順序を、各データストレージシステムを構成しなければならぬ。
 【0010】本発明に係る広域データストレージシステムでは、記憶サブシステムの内部に、冗長化した論理ボリュームを設けることなく、別の記憶サブシステムに対し、データをコピーすることにより上記の問題を解決している。

【0011】また、本発明に係る広域データストレージシステムでは、災害後の復旧作業として、広域ストレージシステムでの再構成を想定しており、正常な運用時に、直接、データ転送を行っていない記憶サブシステム間、で、管理情報を取り取り、データの更新状態を記憶サブシステムで監視し管理する。そして、災害後の復旧作業（再同期、リシンク）において、災害発生直前に各記憶サブシステムが保持しているデータの差分のみを転送することで、即時に、ホストの交代（fail over）と、アプリケーション実行の継続を行う。

【0012】データ更新の順序性を常時保証することについてここで、順序性の保持の時間的範囲について補足説明する。
 【0013】ホストから発行された1/0は記憶サブシステムに書き込まれ、記憶サブシステムが報告するデータの書き込み完了報告を認識し、ホストは次のステップを実行する。ホストは記憶サブシステムのデータ書き込み完了を受領しない場合は障害報告があった場合は、次の1/0を正常には発行しない。従ってデータの書き込みの順序性は、ホストが記憶サブシステムから書き込み完了報告を受領する前後で、記憶サブシステムが順序性を保持の何らかの処理をすることで維持されるべきものである。

【0014】同期転送のリモートコピでは転送されコピーされるデータが付近地又は遠隔地（以下、単に「別地」と略記する。）の記憶サブシステムに書き込まれ、別地の記憶サブシステムからの書き込み完了を受領した後、ホストに対し書き込み完了報告を行う。リモートコピを行わない場合と比較し、リモートコピに係る処理、及びデータ転送処理時間が長くなり、性能が低下する。リモートコピにおける接続距離を延長すると、データ転送に伴う処理時間が増大し、リモートコピを行うことによりホストの1/0処理の性能をさらに低下させる。これを打破する一つの方法が非同期転送である。

【0015】非同期転送は、ホストから1/0を受領した記憶サブシステムが、別地の記憶サブシステムへのデータ転送を行ない別地の記憶サブシステムの書き込み完了を待たずに、ホストから1/0を受領した記憶サブシステムが書き込み完了報告をホストへ返す。これにより、記憶装置サブシステム間のデータ転送は、ホストの1/0処理と関係が無く、ホストの1/0処理と非同期に実行できる。しかし、ホストからのデータの到着順序を守って、別地の記憶サブシステムへデータを書き込まなければ、別地の記憶サブシステムのデータ順序性は維持されず、両記憶サブシステム間でデータの不一致を来す可能性がある。データの順序性を常時保証する機能を追加すれば、このような可能性を極小化できる。
 【0016】別地の記憶サブシステムは、ホスト1/0を受領した記憶サブシステムと比較し、通常はデータの

更新は遅れているが、ホストからのデータ到着順序を守って記憶サブシステムへ書き込み限り、データの順序性に矛盾は無く、ジャーナリングシステムやデータベースリカバリ処理により、障害時の回復が可能である。
 【0017】一方、データの順序性を維持せず、別地の記憶サブシステムへリモートコピしてデータを反映させる方法もある。この方法は、ある時点までのホストから受領したデータを別地へ送り、それらを記憶サブシステムへ書き込む。ある時点までのデータ書き込みが完了する段階で、データ転送を終了し、以降、次の段階まで、リモートコピのデータ転送を中止し、抑止している間のデータ順序性、ホストから受領した1/0の一貫性を保証する。

【0018】この方法では、データの順序性を付与する機能が必要であるが、ある程度の更新分のデータを蓄えておいて、その更新分を一括転送し、リモートへ書き込みが完了した段階で、データ整合性を保証している。この方法ではリモートコピを行っている間に障害が発生すると、リモート側のデータ更新は順序性を維持して更新されていないため全滅となる恐れがある。リモートコピのデータ転送を止めている間のみ、データ整合性を保証でき、adaptiveと呼ばれる。

【0019】出願人の所有する「データの順序性を常時保証する非同期転送によるリモートコピ」の技術によれば、ホストに完了報告を送る際に、記憶サブシステムがデータの順序性を保証する処理をしていることが特徴である。記憶サブシステムの制御装置におけるサーバ側や外部処理の遅延時間内に拘らず、ホストに返す際にデータ順序性をブロック毎に管理する措置を施しているため、常時順序性を保証できる。

【0020】実際には、ホストから受領する1/0発行の間隔よりかなり短い時間で、ブロック毎の制御・管理を行っている。この一方で、リモート側の記憶サブシステムでデータ配置を待たされタイムアウト（Timeout）とする値は、1時間以上に設定可能でもある。大切なのは、出願人のリモートコピの技術が、データに順序性を付与してデータブロックを転送し、これに基づきデータの順序を守って書き込みを行なっている点である。ローカル/リモートのデータ更新の時間差が、例えば半日あっても、順序性を正しければ、更新データ全てを喪失してしまいうる不整合より良いからである。

【0021】
 【課題を解決するための手段】 データを同期及び非同期に転送可能な転送経路、所定の管理情報の取り取りが可能で、データ更新進捗管理手段により、各記憶サブシステムに設けられ、いつ発生するか分からない災害に対応するため、他のデータセンターに設置された記憶サブシステムにおけるデータ更新の進捗状態を、適宜、監視

し、相互にその記憶サブシステムのデータ更新状態を把握させる手段である。

【0023】具体的には、直接データ転送を行っていない記憶サブシステムの各々が、転送状態／ビットマップを持ち、転送ブロックのどの位置が何回更新されたか、一方が問合せ、他方がこれに答えることで、データ更新（リモートコピー）の進捗を監視し管理する機能を有する。

[0024]

【発明の実施の形態】3以上のデータセンタに、それぞれ設置された記憶サブシステムの間を、同期転送により連結する。

【0025】かつ、データの順序性を常時、連続的に保証する非同期リモートコピーの技術で連結する。そして、1箇所のプライマリデータセンターの記憶サブシステムから、これを除いた残りの別拠点の2以上のデータセンターの各記憶サブシステムへ、ホストからプライマリ記憶サブシステムが受領したデータデータを、ホストが更新した順序を保持しつつ、連続的に転送し格納する。

【0026】データが、ホストからの更新順を保証して、元基構造化されるため、万一、データベースに災害、障害が発生しても、残ったデータベース間の差分データのみを転送することで、即時に、リモートコピーの運用構成を回復でき、又は、データ喪失を最小限度とすることができ、

【0027】＜同期・非同期について＞まず始めに、図5、図6を用いて同期伝送によるコピー又は非同期リモートコピーを定義する。

【0028】同期転送によるコピーとは、ホスト1から配電サブシステム1に、データの更新(増込み)指示が有った場合に、その指示対象が附近地に設置された配電サブシステム2にも増込みデータであるとした、配電サブシステム2に対して、指示された更新(増込み)が終了してから、ホストに更新処理の完了を報告する処理手順をいう。ここで、附近地とは、いわゆるメトロポリタンネットワークと称される100km程度までの範囲を言ふものとする。

【0029】つまり、同期転送のリモートコピー（図5）では、ホスト1から受信した更新データブロックを記憶サブシステム1で受信し（■）、そのデータブロックを記憶サブシステム2に転送し（■）、引き込み完了後、これを記憶サブシステム1で受信し（■）、最後にホスト1に対し更新データブロックの引き込み完了を行う（■）。途中の処理に失敗した場合には、ホスト1に引き込み障害を報告する。

【0030】同期転送によるコピーを実施すると、ホスト1に接続された近い方の記憶サブシステム1と、附近に設置された遠い方の記憶サブシステム2のデータの内容が、巨視的にみて常に一致している。このため、災

密により一方がその機能を失った場合であっても、他方の記憶サブシステムに集約直前までの状態が完全に保存されているので、残るシステムで迅速に処理を再開できている効果がある。尚、巨視的にみて常に一致とは、同期化送の機能を実施中には、制御基盤や電子回路の処理時間（μsec, msec）の単位で一貫していない状態が有り得るが、データ更新処理完了の時点ではデータは必ず同一の状態になっていることを意味している。これは、附近地の記憶サブシステムへの更新データの反映が終了しない限り、ホストに近い側の記憶サブシステム1の更新処理を完了できないためである。

【0031】一方、非同期リモートコピー（図6）と
は、ホスト1からこれに接続された近い方の記憶サブシ
ステム1に、データの更新（書き込み）指示が有った場合
その指示対象が遠隔地に位置された記憶サブシステム2
にも書込むデータであって、記憶サブシステム1の更
新処理が終わった後、ホスト1に対し更新処理の完了を
報告し、遠隔地の記憶サブシステム2におけるデータの
更新（反映）が、近い方の記憶サブシステム1におけ
る処理とは非同期に実行される処理手順をいう。

【0032】このため、近江方の記憶サブシステム1で必要とされる処理時間でデータ更新が終了するので、遠隔地の記憶サブシステム2へのデータの格納に起因する伝送時間、格納処理時間等により、ホスト1の処理が待たされることになり、ここで、遠隔地とは、いわゆるトランスコンチネンタルネットワークと称される、附近地より遠いが、距離の制約なく通信又はデータ転送可能な地点を言うものとする。

【0033】より具体的に、非同期リモートコピーでは、ホスト1から受領した更新ブロックを記憶サブシステム1で受領し、ホスト1に対し更新データベースの書き込みを行う。記憶サブシステム1は、自己のスケジュールで、ホスト1の処理とは非同期に、記憶サブシステムへデータを転送する。

【0034】遠隔地又は附近地へのデータ転送経路の複雑化、途中のデータ転送経路のボトルネックにより、データ転送中の当該データの順序性は保証されない（図6、点線の範囲内参照）。

【0035】一般に、データ転送の性能を上げるため、多くは高速転送のため、転送元から複数の転送経路を用いてデータを転送する場合がある。また、転送先まで遠く離れてデータを転送する場合は、転送先から複数の転送経路であっても介在する交換機、ルーターその他の通信用機器により、転送先まで転送経路が1本であることは保証されない。このように複数の転送経路を用いてデータを転送する場合には、経路によっては時間的な差異が生じ、遅い経路と早い経路とを介してデータが送られるため、転送先にあってデータの順序が保存されないのである。

【0036】図6の枠内に一例を示すが、データ転送経路上の順序を、Data#1、Data#2、Data#3、Data#4、Data#5、Data#6の順で示す。

9 - 2003 07 04 17:30

a # 4、Data # 3 としている。記憶サブシステム 2
 における更新順序は Data # 1、Data # 2、Da
 ta # 3、Data # 4 の順序である。記憶サブシステ
 ム 2 において、転送されてきたデータの順序をソートし
 て正規の順序に並べ置いているからである。この更新処
 理の直後に不慮の災害が発生してでも、データ更新の順序
 が守られているため、記憶サブシステム 2 のデータベ
 スやジャーナルファイルシステムは回復処理を行うこと

ができる。逆に、更新処理の直前に災害が発生したとき
は回復処理は不可能であるが、ホストへの応答とは非同
期に、記憶サブシステム間で逆流的に間断無くデータ転
送処理を行うことで、データ不整合を極小化でき、巨視
的に見て、常時、更新データの順序性を確保できる。

【0037】本実施の態様では、ホスト1からデータテーブルを受領し、記憶サブシステム2へ転送する際に、ホスト1からのデータ更新順序を示すシーケンス番号情報（シーケンス番号）をデータに付して転送している。このため、記憶サブシステム2で、シーケンス番号情報に基づくソート制御を行い、順序性を保証して、データの格納を完了できる。

このような一連のデータ転送・処理に必要な処理時間の後は、データの順特性が遠隔地の記憶サブシステム2において保持されている。このように非同期的コピーを、これに固有なデータ処理を連続して行うこと（非同期的モートコピー）で、常時、データ更新の順序性を保証することができる。

【0038】非同期リモートコピは、ホスト1の処理性能を落とさず、記憶サブシステム1及び2の間の距離性を拡大できる長短があり、かつ、常時、データの順次性が保証されるため、広域データストレージシステムの利用者が自己の業務を遂行する上で、ほぼ任意の時点でのデータベースやジャーナルファイルシステムの整合性を、遠隔地に設置された記憶サブシステムにおいて確保できる長短を有している。

【0039】＜広域データストレージシステム、その1＞図1に本発明の広域データストレージシステムの全体構成を示す。図9は本発明の別の広域データストレージシステムの全体構成を示す図である。図10は図1と図9の構成の組み合わせによる応用例を示す図である。

【0040】図1において3つのデータセンタに記号サブシステムを設置する。各データセンタには複数の記憶サブシステムが設置されるも良いし、それらがリモートコピー機能に伴った接続形態となっても良い。アプリケーションはデータセンタ1に接続されたホストで実行される。尚、ホストとデータセンタ1とのデータ転送経路は、ファイバーチャネル、メインフレームインタフェース、イーサネット（登録商標）LAN、公衆回線、インターネットその他の通信回線である。

【0041】データセンタ1とデータセンタ2は近接地に存在し、同期転送によりデータ転送し得る構成である。データセンタ1とデータセンタ3は遠隔地に存在

し、これらの間は非同期リモートコピーの技術によりデータ転送し得る構成である。

【0042】正常な運用形態では、ホストからデータベース1が受領した更新データは、データベース1に登録センターに登録される。この更新された登録サブシステムに格納されたデータセンター2の記憶装置にデータを、附近地に設置されたデータセンター2の記憶装置に格納する。サブシステムへ、ファイバーチャネル、メインフレーム、インターネットLAN、公衆回線、インフラファース、イーサネットLAN、公衆回線、インターネットその他専用回線を介して、同様に転送される。

つまり、データベース1とデータベース2では、記憶装置間のデータの整合性は恒時的には保たず保たれている。

【0043】 正常な運用形態では、また、ホストからデータセンタ1が受領した上記の更新データを、遠隔地に設置されたデータセンタ3の記憶体サブシステムへ、上述と同様にデータセンタ3の記憶体サブシステムと同期化と同様な専用回線を介して、上記の同期化処理と同様に、非同期リモートコピーの技術で転送される。尚、データセンタ1とデータセンタ2、データセンタ1とデータセンタ3、それぞれの間のデータ転送経路は同一種類の回線にする必要はない。

【0044】データセンタ1とデータセンタ3との間は遅延距離であり、この間の転送に起因する更新データの到着順序の不整合が生じる。また、転送元となるデータセンタ1の記憶サブシステムには、転送先で未放映のデータとなる差分データが、保存することとなる。しかし、

本発明の非同期リモートコピーでは、所定の非同期処理中に固有のデータのデータ処理後は、データベースやファイルシステムのデータの回復処理に必要な、ホストからのデータ順序性を保証しているために、不整合を生じたいデータを確認してデータを回復させることが可能である。この結果、データベース1とデータベース3の記憶サブシステム4では、上記ホストから受領した更新データの順序性は保たれる。

【0045】データセンタ2とデータセンタ3の間は、万一のリカバリ処理に備え、データを転送する通信経路は、複製・準備されているが、このストレージシステムは、複製・準備時にホストからの更新データは転送されない。データセンタ1での災害・障害発生の際に備え、正、常な運用形態で、データ転送の進捗状態を知らせるコマンドが、この通信経路を介して、データセンタ2からデータセンタ3へ、又は逆にデータセンタ3からデータセンタ2へ、送受を置けることとなる。尚、複製・準備された通信経路は、ファイルチャネル、メインフレーム、インタフエース、イーサネットLAN、公衆回線、インターネットその他専用回線である。

【0046】正常時には、記憶サブシステム1と記憶サブシステム3との間の非同期リモートコピーにより与えられたホストからの更新データの到着を、記憶サブシステム2から発生する“データ進捗問合せコマンド”により、データセンタ2とデータセンタ3の間の通信線路を介して、問合せる。

2003 07 04 17:30

【0047】“データ進捗問合せコマンド”の起動は、記憶サブシステム2のスケジュールに従って為される。記憶サブシステム1からの同期転送によるデータの受領のタイミングで、当該コマンドを発行しても良いし、所定の時間間隔で繰り返して問合せも良い。所定の時間間隔としては、例えば10msecから500msec毎に問合せても良いが、後述する転送状態/ビットマップの管理、これに基づく差込データの管理に時間がかかり過ぎない程度となる。尚、1回の問合せで複数のビットマップを検査するようにしても良い。

【0048】正常な運用時には、記憶サブシステム2と記憶サブシステム3との間で、直接、データの転送は行なわれない。このため、記憶サブシステム2が“データ進捗問合せコマンド”を発行して、記憶サブシステム1と記憶サブシステム3のデータ更新状況を把握する。

【0049】万一、データセンタ1で障害が発生したときには、データセンタ2のホストを用いて、これまでのシステム運用を続行し（ホストのフェールオーバー）、記憶サブシステム2と記憶サブシステム3との間の差分データを、リカバリ処理に備えて敷設されたデータ転送の通信線路を用いてデータセンタ2からデータセンタ3へ転送する。差分データのみを転送で即時に広域データストレージシステムを回復させることが可能である。尚、フェールオーバーとは、プライマリシステムからサブシステムへ切り替えることをいい、古くは、ホットスタンバイとも呼ばれていた。

【0050】この後、データセンタ2からデータセンタ3へ、上記の通信線路を用いて、上述のような非同期リモートコピーを行うこととすれば、データセンタ1の復旧に伴う、データセンタ2とデータセンタ1との間の同期転送の復旧により、障害発生前の広域データストレージシステムを復旧させることができる。但し、障害発生前後で、データセンタ1とデータセンタ2の役割が入れ替わっている。

【0051】このように、近傍地に存在する2つのデータセンタと、遠隔地に存在する2つのデータセンタとを統合し合計3つのデータセンタとすることで、リモートコピーの技術で連結する広域データストレージシステムとする。こうしておけば、中小規模の災害・障害のときは、近傍地に存在する、相互に同期転送により連結されたデータセンタの一方で他方の代替を行うことができ、2つのデータセンタの記憶サブシステムのデータは同期転送により巨额的にみて一致しており、フェールオーバーが即時に行なえるからである。

【0052】＜広域データストレージシステム、その2＞図1のデータセンタ2とデータセンタ3との間の通信線路が非常用であるため、この通信線路を選択せず、障害・災害復旧後のデータセンタ1とデータセンタ3との間のデータ転送線路を選択する場合には、復旧後は、広域データストレージシステムは、図9の構成となる。

【0053】図9は、記憶サブシステム1と記憶サブシステム2が同期転送で、記憶サブシステム2と記憶サブシステム3が非同期リモートコピーで、それぞれ、接続された例である。図1の広域データストレージシステムにおいて、データセンタ1からデータセンタ2へ運用を切り替え、データセンタ2を主たる運用サイトとし、災害・障害復旧後は、データセンタ2からデータセンタ1へデータセンタ2を同期転送させる一方で、データセンタ1からデータセンタ3へデータを非同期転送させる構成となるからである。

【0054】図9の場合において、直接データ転送に關与しない記憶サブシステム1から記憶サブシステム3へ、“データ転送進捗問い合わせ”コマンドが発行され、データセンタ3が応答して結果をデータセンタ1へ返す構成となっている。また図10は、図1と図9を組み合わせた構成である。記憶サブシステム3と5との間、記憶サブシステム2と5との間が、“データ進捗問い合わせコマンド”の発行・応答の経路に該当する。

【0055】上記の広域データストレージシステムの態様であれば、大規模な災害や、近傍地に存在する2つのデータセンタに相次いで障害が発生した場合であっても、データセンタ3のホストへフェールオーバーすることと、災害直前のシステムが運用してきたデータを引き継いで処理でき、また、データの喪失を最小限度とすることができ。

【0056】つまり、近傍地にある2つのデータセンタが全滅する程度の災害が発生したときは、遠隔地に存在するデータセンタ3又は5（図1、図9、図10）の記憶サブシステムを生かすことができる。ホストからの更新データの順序性が確保されつつ、非同期リモートコピーが行なわれているからである。但し、災害による未反映のデータは復旧できない。

【0057】＜記憶サブシステムの構成について＞図1、図9及び図10では、同期転送によるコピー及び非同期リモートコピーの組み合わせを示している。本来、リモートコピーは、1論理ボリュームと1論理ボリュームをデータ転送技術で結合したものである。本発明では、1個の論理ボリュームに対するデータ受領を、同期転送し、更に非同期転送して、近傍地と遠隔地の双方のリモートコピー機能でデータ送信制御を行なっている。

【0058】これらは記憶サブシステムの制御装置のマイクロコードで実現される機能である。ホストや別の記憶サブシステムからの更新データは、一旦、キャッシュ5（図2）に格納される。この時点では、当該データは、まだ記憶サブシステム内のハードディスクドライブにRAID制御により書き込まれていない。キャッシュ5内で当該データの転送制御情報を加え、別の記憶サブシステムへリモートコピー転送したり、複数の記憶サブシステムとのリモートコピー構成を同時に実現する制御を行う。同期転送と非同期転送による組合せを守ること

により、いつ災害が発生しても、各データセンタでは、データの更新順序を保った、データベースやジャーナルファイルシステムがリカバリ可能な論理ボリュームを常時、保持していることとなる。

【0059】図2は、記憶サブシステムの構成構成を示す図である。

【0060】制御装置1は、ホスト及びリモートコピーの接続先とデータの送受を行うチャネルアダプタ3、ディスク装置2内のハードディスクドライブ7をディスクインタフェース8（ディスク1/F8）を介して制御するディスクアダプタ9を有する。

【0061】チャネルアダプタ3とディスクアダプタ9は、それぞれ、マイクロプロセッサを有し、データ転送パス1・制御パス11を介してキャッシュメモリ5とバス11・制御バス11を介してキャッシュメモリ5と接続されている。尚、バス構成は一例であり、必要に応じてクロスバ構成としても良い。また、制御装置1を複数設けてクラスタ構成とし、複数の制御装置1を連結する共通の第3のバスを追加しても良い。

【0062】ホストとの間や、リモートコピーの接続先とデータ送受を行う際の格納元は、キャッシュ5であり、制御装置1、構成管理情報、転送状態/ビットマップは、制御メモリ6に格納されている。

【0063】リモートコピーには送信及び受信の機能が有り、本実施例ではホストからI/Oを受領するチャネルアダプタを分けて搭載している。ホストから受領したI/Oは、一旦、キャッシュ5へ格納される。リモートコピーの転送先情報や後述する状態管理/ビットマップは、制御データとて制御メモリ6に格納され、マイクロコードにより制御される。

【0064】キャッシュに受領したデータは、ディスクアダプタ9によりハードディスクドライブ7へRAID制御で書き込まれる。これは別の処理である、マイクロコードを用いた制御により、予め定義されたリモートコピー転送先への送信制御が行なわれる。

【0065】例えば、ホストから受領したデータが、後述するリモートコピーの対象であり、非同期転送によるデータ送信を行うと定義されていた場合には、キャッシュ5の内部のデータに対して、データ受領順にシーケンス番号を付与する。これはデータ更新を示すID情報でもある。シーケンス番号を付与されたデータは、チャネルアダプタ3のリモートコピー送信機能により、当該シーケンス番号と共に送信される。

【0066】別の実施例で、ホストから受領した更新ブロックを、複数の論理ボリュームと接続するリモートコピー制御が定義されていた場合には、キャッシュメモリ5の内部のデータは、同期転送用に加工されると同時に、非同期転送用にも加工され、シーケンス番号が付与されて、それぞれ、チャネルアダプタ3で近傍地又は、遠隔地に向けて送信される。

【0067】図2は本発明を実現する一例であり、本発

明はハードウェア構成に依存しない。リモートコピー接続が記憶サブシステム間で実現可能であれば、マイクロプロセッサによる論理的なサポート、マイクロコード制御で実現できるためである。

【0068】＜転送状態/ビットマップ＞図4は、転送状態/ビットマップ（適宜、ビットマップと略記する。）の一例を示したものである。これは、直接データ転送を行っていない2つのデータセンタに設置された記憶サブシステム2の内部に、災害・障害の復旧の際に、ペアを組むであろう相手（別のデータセンタ）に設置された記憶サブシステム）のデータ更新の進捗状態を知るために用意されたものである。例えば、図1ではデータセンタ2とデータセンタ3との間で、非常時のためにペアが組まれる。図9の広域データストレージシステムであれば、記憶サブシステム1と記憶サブシステム3との間で、図10では、記憶サブシステム2と記憶サブシステム5、記憶サブシステム3と記憶サブシステム5との間で、それぞれ、非常時のために、ペアが組まれることとなる。

【0069】転送状態/ビットマップは、ペア（対）となる論理ボリュームに対して必要であり、本発明では1個の論理ボリュームの実体に対し、2個以上の転送状態/ビットマップを持ち得る。各ビットマップは、ペアやペアとなる場合を想定した定義付けにより、相手の論理ボリュームとの差分管理を行うために使われる。ビットマップの中のブロックナンバは、論理ボリュームの更新を管理する最小単位であるブロックに対応させた番号である。

【0070】ホストI/Oは、このブロックナンバと同一単位である必要はない。ホストI/Oの単位は、通常、最小で512バイトとされ上限も設けられているが、可変である。一方、ビットマップは、50kBの大きさ、又は700kB程度の大きさのものもあるが、20kBから1000kB程度まで種々の大きさがある。ホストI/Oの1ブロックに対して、必ずしも、1ビットマップが対応する訳ではない。

【0071】ブロックナンバに対応するブロックの内容が更新されれば、差分管理は当該ブロックナンバ全体となり、同期（リシंक）を行うときに当該ブロックナンバのデータ全体が転送されることとなる。

【0072】ビットマップは、ブロックナンバ毎に、当該制御ボリュームの更新された単位として、リモートコピーによるペアを再構築する際（再同期、リシंक）に当該更新されたブロックのみを転送する目的で、相手論理ボリュームに転送すべきUpdate情報を持つ。つまりUpdateフラグがOn（図4の実施例では1）、であれば転送対象であることを示す。通常のUpdateは、ホストからのコマンド単位で為されることから、カウンタ値が0であることに基づき、Updateフラグを0とする。

【0073】ビットマップは更に、同一ブロックナンバで、複数回の更新を記録するカウンタ値を持つ。カウンタ値は、更新が無ければ「0」、3回更新されれば「3」となる。ブロックナンバで表されるデータブロックの大きさが、ホストから更新されるデータブロックより大きい場合には、このカウンタ値を使うことにより領域に相手論理ボリュームへ更新データのみを転送できることとなる。

【0074】後述の「データ進捗問合せコマンド」の中に格納されたブロックナンバとカウンタ値と、同優先の記憶サブシステムのビットマップのブロックナンバとカウンタ値との比較を、データコピ監視機能（後述）で行う。この際に、ある記憶サブシステムが持つカウンタ値が、この記憶サブシステムに送付されて来た「データ進捗問合せコマンド」に記述されたカウンタ値と等しいか大きい場合に、所定の記憶サブシステムのビットマップのカウンタ値は1減算される処理を受ける。

【0075】送付されて来た「データ進捗問合せコマンド」に記述されたカウンタ値未満である場合は、その記憶サブシステムのビットマップのカウンタ値は何ら処理を受けない。そして減算したくないかを、「データ進捗問合せコマンド」に格納して返す。

【0076】その記憶サブシステムのビットマップのカウンタ値が、送付されて来た「データ進捗問合せコマンド」に記述されたカウンタ値「以上」の場合には、データ更新の進捗は、正常なリモートコピ機能により既に、その記憶サブシステムにおいて格納済み、書き込み済みであることを意味する。また「未済」の場合には、データが未到着であることを意味している。

【0077】図4のカウンタ値は有限であり、例えば、1バイト分をカウンタ値として割り当てた場合には、256回を超え管理はできない。この例では同一ブロックが256回を超えた更新を受けた場合には、最早、カウンタ値のUpを行わず、Updateフラグを恒久的に立ててしまう処理を行う。つまり図4でカウンタ値に「Over Flow」を意味する情報を格納する。

【0078】このような恒久的な指定がなされると（図4、Over Flow）、ビットマップで特定される、恒久的指定のなされたブロックのUpdateフラグの解除（0を入力すること）は、相手論理ボリュームへの転送が完了コピが確定したことを、このビットマップを有する記憶サブシステムが認識するとまで行なわれない。

【0079】カウンタ値を用いた更新管理を行う理由を次に補足説明する。

【0080】例えば、50k B程度のデータ直を有するブロックに対応させてビットマップの管理を行う場合、この50k Bのデータのうち、異なる3箇所が、異なる時刻において、それぞれ更新されたところ、トラッキングに対応させてビットマップ管理を行うのは、災害・障

害後の復旧（再同期、リシンク）において扱う単位がトラック単位であるためである。

【0081】カウンタ値による管理を行なわない場合には、Updateフラグのみ監視することとなるが、ある時刻でUpdateフラグが1であることのみを認識しても、その後の時刻に2度目、3度目の更新があった場合には、2度目以降のデータ更新を見逃してしまう。新たにカウンタ値の概念を導入して、ホストからコマンド単位で与えられるデータブロック（ここではトラックの一部）の更新を微細に監視することで、かかる不都合を防ぐことができる。

【0082】次に、図2の制御装置1の内部でマイクロコードにより実現される転送状態/ビットマップの機能について定義する。論理ボリュームはリモートコピの対となる論理ボリュームとの間で下記の転送状態を有する。これらは同期転送又は非同同期転送に依存しない。

【0083】1) 「正常ペア状態」とは、データの順序性を保証して、双方のボリューム間で、同一のデータを2重に保持している状態をいう。

【0084】2) 「転送抑制ビットマップ登録の状態」とは、データの更新をビットマップに登録する状態をいう。未だペアの相手ヘデータの転送は行なわれていない。

【0085】3) 「ビットマップ使用のコピー状態」とは、「転送抑制ビットマップ登録の状態」から「正常ペア状態」への移行期をいう。2重化のためのコピの初期状態に当たる。

【0086】4) 「障害状態」とは、障害によりデータを転送できない状態をいう。ビットマップに登録される。

【0087】5) 「ペア無ビットマップ登録状態」とは、本装置固有の特殊な状態をいう。災害・障害前、相互にデータ更新状態を監視保持する必要から生じた状態である。

【0088】6) 「ペア無状態」とは、ビットマップは用意されているが、未だペアを組んでおらず、データ更新の情報が登録されていない状態をいう。

【0089】「ペア無ビットマップ登録状態」が存在することが本装置の特性となる。この状態を持つことなると、「転送抑制ビットマップ登録の状態」というサスペンド（Suspend）状態が兼ねてもあり、ここで、サスペンド状態とは、論理ボリュームへのデータの更新状態を、ビットマップでのみ管理し、リモートコピによる転送制御を行なわれない状態をいう。

【0090】「ペア無ビットマップ登録状態」を持つのは、転送状態/ビットマップをペアで持つ必要からである（図3）。例えば、図1の広域データストレージシステムにおいては次の理由による。

【0091】データセンター3が保持するデータの監視するため、データセンター2の記憶サブシステムの論

理ボリュームに対処して設けられた転送状態/ビットマップに、データセンター3のデータ更新状態を持つ必要があり、且つ、データセンター2が保持するデータを監視するため、データセンター3の記憶サブシステムの内部の論理ボリュームに対処して設けられた転送状態/ビットマップに、データセンター2のデータ更新状態を持つ必要があるためである。

【0092】図9の広域データストレージシステムにおいては、データセンター2の障害発生に備えて、データセンター1とデータセンター3のリモートコピの差分管理情報から、データセンター1とデータセンター3との間でペア構築を目的として、「ペア無ビットマップ登録状態」をデータセンター1とデータセンター3で持つ必要がある。この結果、記憶サブシステムやデータ転送経路のどこに障害が発生しても、状態把握が可能で、ビットマップによる未転送データブロックの記憶と、障害回復後に更新部分のみの差分転送が可能となる。

【0093】転送状態/ビットマップの機能は、上記の様な制御を実現するマイクロコード及びビットマップと関連する制御テーブルから成る。具体的機能は、例えば、図2のマイクロセッサ4のマイクロコードと制御セッサ6で行なわれるが、先に示した様にマイクロコードの制御により自由に実装できる。例えば、マイクロセッサ10による実装も可能である。またマイクロプロセッサが1台のみの制御装置でも実現できる。

【0094】＜広域データストレージシステムの運用＞図3は、図1の広域データストレージシステムが正常に運用されている場合の基本的制御方法を説明するため

のブロック図である。正常運転ではデータ進捗問合せコマンドを記憶サブシステム2から記憶サブシステム3へ送信する。例えば、記憶サブシステム1の障害の際、実際の差分データの転送に際しては、記憶サブシステム2と記憶サブシステム3との間で、転送状態/ビットマップの機能を使用し、両方の記憶サブシステムのビットマップについて、論理演算を行う。その結果に基づき、相当するデータブロックのみを記憶サブシステム2から記憶サブシステム3へ転送している。図8に、図1の広域データストレージシステムのデータセンター1に障害・災害が発生した場合において、非同同期リモートコピを再開させる経路の手順を示す。

【0095】図8において、正常な運用では、データセンター1から近隣のデータセンター2へ同期転送によりデータの二重化が図られる一方で、遠隔地のデータセンター3へは非同同期転送によりデータの更新順序を確保したコピーが行なわれている。そして、データセンター2の記憶サブシステム2のスケジュールで、データ進捗問合せコマンドがデータセンター3に對し発行され、データセンター2と3とは管理情報をやり取りして、データの差分管理を行なっている。

【0096】データセンター1に災害・障害が発生する

と、データセンター2の記憶サブシステムは、非同同期転送により、差分データをデータセンター3へ送付し、即時に、データセンター2と遠隔地のデータセンター3によるシステム運用を回復できる。

【0097】図3において、転送状態/ビットマップは、1論理ボリューム当たり2個持ち、それぞれが、これらのビットマップを用いた機能を実行する。記憶サブシステム1は、記憶サブシステム2と記憶サブシステム3に

対し、転送状態/ビットマップ#1に対応する機能及びビットマップ#2に対応する機能を持つ。

【0098】記憶サブシステム2と記憶サブシステム3は、同期転送及び非同同期転送の各々について転送状態/ビットマップ#3及び#6の機能をそれぞれ持つ。これは#1と#3、#2と#6のそれぞれの機能は、正常運転の際には、「正常ペア状態」を格納している。

【0099】転送状態/ビットマップ#4及び#5の機能は、それぞれ、記憶サブシステム2及び記憶サブシステム3が持つている。この広域データストレージシステムが正常に運用されているときには、転送状態/ビットマップ#4及び#5の機能は、上述の「ペア無ビットマップ登録」状態を保持する。

【0100】転送状態/ビットマップ#4の機能は、記憶サブシステム3の論理ボリュームに対する差分管理を、それぞれシステム2の論理ボリュームに対する差分管理を、それぞれ行う。

【0101】図10の拡張として、ホストからのI/Oを受領する、第1のデータセンターに設けられた記憶サブシステムの制御装置1が、N台の非同同期転送のコピー先と、M台の非同同期リモートコピのコピー先を持つ構成では、その制御装置1は、N+M個の転送状態/ビットマップの機能を有する。これに対応する、遠隔地又は近地の記憶サブシステム（コピ先）も、転送状態/ビットマップを持つこととなる。この結果、制御装置1やデータ転送経路のどこに障害が発生しても、状態把握が可能で、ビットマップによる未転送データブロックの記憶と、災害回復の際の更新部分のみの差分転送が可能となる。

【0102】＜データコピ監視機能＞次に、データコピ監視機能について説明する。この機能には、ビットマップの制御機能、リモートコピのステータス管理機能、繰越管理機能、データ進捗問合せコマンドの制御機能、リモートコピのデータ転送指示機能等が含まれる。

【0103】図3の記憶サブシステム2の制御装置で、同期転送によるデータブロックを記憶サブシステム1から受領する。本データは記憶サブシステム2のキャッシュメモリに格納されディスクドライブで記憶される。この際、転送状態/ビットマップ#4の機能により当該データブロックが登録される。図4のビットマップに登録

する。

【0104】次に当該ブロックナンバとカウンタ値を格納した"データ進歩同合せ"コマンドを記憶サブシステム2から記憶サブシステム3に対して発行する。発行のタイミングは同期転送に基づいても良いし、記憶サブシステム2の独自のスケジュールで行なっても良い。

【0105】記憶サブシステム3の制御装置で、記憶サブシステム2からの"データ進歩同合せ"コマンドを受領し、転送状態/ビットマップ#4のブロックナンバとカウンタ値を切り出し、記憶サブシステム3の該当する転送状態/ビットマップ#5のそれらと比較する。

【0106】その結果、転送状態/ビットマップ#5のブロックナンバがUpdateフラグ1(更新)を示し、かつ、カウンタ値が転送されて来たもの以上であれば、同期転送に係るデータと、非同期リモートコピに係るデータとが一致しているため、転送状態/ビットマップ#6の対応するブロックナンバから、カウンタ値を1減算する。

【0107】減算の結果、カウンタ値が"0"となった場合には、Updateフラグを"0"とする。カウンタ値が"Over Flow"である場合には、何も操作しない。

【0108】また、転送状態/ビットマップ#5に登録されていたカウンタ値が、記憶サブシステム2からの同合せコマンドから抽出されたカウンタ値未満であったら、Updateフラグが"0"(Of)で更新が示されなかった場合には、#5への更新は行なわず、これをデータ進歩同合せコマンドの結果として記憶サブシステム2へ返す。

【0109】#5の転送状態/ビットマップの機能が、#6の転送状態/ビットマップのカウンタ値を減算するということは、記憶サブシステム1から同期転送により既に記憶サブシステム2へ到着したデータブロックが、記憶サブシステム1から記憶サブシステム3へ非同期転送により到着済であったことを意味している。

【0110】データコピ監視機能は、本装置結果を用いて記憶サブシステム2の転送状態/ビットマップ機能に問合せ"コマンドのブロックナンバとカウンタ値が既に登録されていた旨の応答を返す場合(減算できた場合)には、記憶サブシステム2の制御装置でも転送状態/ビットマップの機能でカウンタ値の減算、Updateフラグの操作を同様に行う。

【0111】当該コマンドの応答結果が、未登録であれば、記憶サブシステム1から記憶サブシステム3へのデータの非同期転送が未完であるとして、記憶サブシステム2の転送状態/ビットマップ#4の機能は、自己のビットマップに更新状況を保持する。これは後に更新差分部分のみを再同期させる際の対象となる。

【0112】この時点で記憶サブシステム1が重大障害

を持ち、記憶サブシステム2と記憶サブシステム3との間でリモートコピ構成を再構築(再同期、リシンク)しなければならぬ場合には、ビットマップを参照した結果、未転送のデータのみ、即ち、差分のデータブロックのみを、記憶サブシステム2から3へ転送すれば良い。その結果、差分データの転送だけで即時に"正常ベア"を構築できる。これを実現する機能を"データコピ監視機能"と呼ぶ。

【0113】<正常な運用の際に、直接データ転送を行なわない記憶サブシステム間での差分管理方法、その1>図9の広域データストレージシステムにおいて、記憶サブシステム2に障害が発生した場合に、記憶サブシステム1と記憶サブシステム3との間で非同期リモートコピによるシステム運用の復旧を図るときを考慮する。

【0114】このために、ホストからデータ更新を受領した記憶サブシステム1(図2)は、記憶サブシステム2の制御装置1の論理ボリュームに同期転送のコピーによるデータ転送を行う際に次の処理を行なう。

【0115】転送するブロックの位置情報を、記憶サブシステム1の制御装置1に存在するビットマップに、記憶サブシステム3の論理ボリュームの更新情報を格納する。このとき既に転送したブロックが記憶サブシステム3において更新されていたときは、ビットマップのカウント値を1増加(インクリメント)する。

【0116】記憶サブシステム1の制御装置1は、記憶サブシステム2の制御装置1に対して同期転送が完了した後、記憶サブシステム3の制御装置1に対して、同期転送したデータブロックが、記憶サブシステム2の制御装置1を經由して到着したか否かを問合せため、記憶サブシステム1と記憶サブシステム3とを結ぶ通信線路を用いて、確認コマンドを発行する。

【0117】確認コマンドには、ホストから受領した更新データの記憶サブシステム1におけるデータブロックのブロックナンバとカウンタ値が含まれている。確認コマンドを受領した記憶サブシステム3の制御装置1は、記憶サブシステム2の制御装置1経由で既に存在するデータブロックが、確認コマンドで同合されたブロックと一致するか否かを判定する。

【0118】記憶サブシステム3の制御装置1は、記憶サブシステム2の制御装置1の論理ボリュームに対する転送状態/ビットマップの機能の他に、記憶サブシステム1の制御装置1の論理ボリュームに対する状態管理/ビットマップの機能を持つ。

【0119】記憶サブシステム3の制御装置1は、記憶サブシステム2の制御装置1からデータを受領すると、記憶サブシステム1の制御装置1の状態を格納すべく、自己の持つ転送状態/ビットマップへ登録する。このビットマップでは、論理ボリューム内のアドレスに関わるブロック位置に対する更新情報を有し、さらに複数回の同

期ブロックに対する更新を管理するためにカウンタ値を有している。

【0120】記憶サブシステム3の制御装置1の転送状態/ビットマップに登録された結果は、記憶サブシステム1の制御装置1から発行された確認コマンドのブロックナンバ及びカウンタ値と比較される。比較の結果、一致又はカウンタ値が確認コマンドにあったカウンタ値以上の場合には、データの到着が正常に完了していると判断し、転送状態/ビットマップの機能を用いてビットマップのカウントを1減算する。

【0121】他方、記憶サブシステム1の制御装置1は、記憶サブシステム3の制御装置1から返される結果が、記憶サブシステム3ヘデータブロックが記憶サブシステム2を経由して到着していることを示す場合には、上述の記憶サブシステム3の制御装置が為したように、転送状態/ビットマップの機能を用いてビットマップのカウントを1減算する。

【0122】以上のように、ビットマップを監視・管理することで、記憶サブシステム2が災害等により重大障害を持ち、同期及び非同期転送によるデータ送受が行なえなくなってしまう場合であっても、ホストがI/Oを発行する記憶サブシステム1と、記憶サブシステム2の内容を非同期リモートコピにより格納した記憶サブシステム3との間で、非同期リモートコピを構成することができ

る。

【0123】この際に記憶サブシステム1と3の、それぞれの制御装置の転送状態/ビットマップの機能により、論理ボリュームの全データをコピすることなく、差分データのブロックのみを転送することにより、即時に、構築することができ

る。

【0124】<正常な運用の際に、直接データ転送を行なわない記憶サブシステム間での差分管理方法、その2>図10の広域データストレージシステムにおいて、ペアテム1と2並びに記憶サブシステム1と3、それぞれの間のデータ更新状態の管理のために、転送状態/ビットマップの機能が各論理ボリューム毎に用意される。

【0125】記憶サブシステム1の制御装置1で障害が発生し、同期転送のコピー及び非同期リモートコピの双方が継続不可能となった場合には、記憶サブシステム2と3のそれぞれの制御装置1の間で、先ず差分データをコピし両者を一致させる。次いで、記憶サブシステム2と3の間に非同期リモートコピを構成する。

【0126】ホストから更新すべきデータを受領した記憶サブシステム1の制御装置1は、記憶サブシステム2の制御装置1へ同期転送によりデータブロックを送出し、これを記憶サブシステム2の制御装置1を受領する。記憶サブシステム2の制御装置1は、受領したデータブロックの位置情報(ブロックナンバ)を、記憶サブシステム3の制御装置1の配下の論理ボリュームの管理

情報との比較のために、自己が保持する転送状態/ビットマップに格納する。転送状態/ビットマップは、受領したデータブロックが更新された場合には、カウンタ値を1増加(インクリメント)する機能を備え、複数回のデータブロックの更新を記録できる。

【0127】記憶サブシステム2の制御装置1は、上記の転送状態/ビットマップへ所定の管理情報を登録した後、記憶サブシステム2の制御装置1と記憶サブシステム3の制御装置1との間を結ぶデータ転送線路を用いて、データブロックが記憶サブシステム3へ到着したか否かを問合せ確認コマンドを、記憶サブシステム3の制御装置1へ発行する。

【0128】確認コマンドは、記憶サブシステム2の制御装置1が、同期転送により記憶サブシステム1から受領したデータブロックの位置情報であるブロックナンバと、データブロックが何回更新されたかを示すカウンタ値を含む。

【0129】記憶サブシステム3の制御装置1は、記憶サブシステム1の制御装置1から、非同期リモートコピの技術で受領したデータブロックの位置情報(ブロックナンバ)とカウンタ値を、記憶サブシステム3の制御装置1の配下の論理ボリュームの管理情報としての比較のために、自己の制御装置1が持つ転送状態/ビットマップの機能を用いてビットマップに格納する。記憶サブシステム3の制御装置1は、ビットマップと確認コマンドの、有する対応する値との比較を行う。

【0130】記憶サブシステム2から3へ問合せた確認コマンドが有するブロックナンバとカウンタ値と、記憶サブシステム3の制御装置1が持つ、記憶サブシステム2の制御装置1の配下の論理ボリュームの管理情報である、これらの値とを比較して、確認コマンドの値と同一又はカウンタ値が確認コマンドのカウント値より大きい場合には、転送状態/ビットマップの機能で、そのデータブロックのカウント値を1減算する。

【0131】減算した結果が0になる場合は、記憶サブシステム2と3との差分データはないことになるので、ビットマップの管理から削除する。上記の比較の結果が一致しない場合には、記憶サブシステム3の制御装置1は、ビットマップのカウント値と操作しない。

【0132】記憶サブシステム3の制御装置1は、記憶サブシステム2の制御装置1に、確認コマンドの応答である判定結果を返す。この結果を記憶サブシステム2の制御装置1が参照し、比較がカウンタ値を減算した場合には、既に記憶サブシステム2と3の間で、同一のデータブロックの更新が正常に終了していると断定する。

【0133】記憶サブシステム3に更新すべきデータブロックが届いていない場合には、記憶サブシステム2のみ、更新に係るデータブロックが格納されていることになる。記憶サブシステム2の制御装置1は、自己の転送状態/ビットマップの機能でこれを記録する。

【0134】記憶サブシステム2の制御装置1が、記憶サブシステム3の制御装置1から確認コマンドの応答を受領し、記憶サブシステム3に更新すべきデータブロックが未到着であった場合には、記憶サブシステム2の制御装置1が待つ、記憶サブシステム3の論理ボリュームの更新状態に対応する転送状態/ビットマップのカウント値は減算しない、このことは、そのビットマップは、更新に係るデータブロックが、記憶サブシステム2と3との間で差分であることを示す。

【0135】他方、データの到着完了を示した場合に、上記の転送状態/ビットマップの更新に係るデータブロックのカウント値を1減算する。カウント値が0のときは、記憶サブシステム2と3との間で、更新に係るデータブロックは同一であり不整合がないので、差分データのコピーの対象とはしない。

【0136】このように、正常運用の際に、直接データ転送を行っていない記憶サブシステムの制御装置同士が、災害・障害からの回復を想定して、論理ボリュームの差分データ管理を行っているため、記憶サブシステム間で差分データのみをコピーし不整合をなくすることが高進に行なえる。

【0137】＜フェールオーバー後のシステムの運用＞
図7に、図1の広域データストレージシステムが、フェールオーバーにより状態を遷移して図9の構成となった場合の運用について簡単に説明する。図3で記憶サブシステム1に、図9で記憶サブシステム2に、図10で記憶サブシステム1、2又は4に、それぞれ重大な障害が起きた場合には、図7に示す様に、残存する2以上の記憶サブシステムの間で、リモートコピー構成の復権を図ることとなる。

【0138】本発明によれば、図7の様に、直接データ転送に関係しなかった論理ボリューム間（記憶サブシステム1と記憶サブシステム3との間）で、差分データのみのコピーすれば、即時に、リモートコピーのペアを生成でき、リモートコピーの運用再開が可能である。

【0139】本発明を実施しない場合には、図3の記憶サブシステム2と3の間、図9の記憶サブシステム1と3の間で、それぞれ、リモートコピー構成をつくるに際し、図3の構成では記憶サブシステム2から記憶サブシステム3に対して、図9の構成では、記憶サブシステム1から記憶サブシステム3に対し、それぞれ、記憶サブシステムが保持するデータのフルコピーを行なわなければならない。大規模のデータセンターでは、コピーに長時間を要し、リモートコピーの運用再開が遅くなる。長時間を要するコピー中に、再度、コピー元やデータ転送経路に障害・故障が発生すると、データは破壊され喪失することとなる。

【0140】図11を用いて、図9の構成におけるデータコピー監視機能について簡単に説明する。

【0141】データ進捗問合せコマンドは記憶サブシ

ステム1から記憶サブシステム3に対して発行される。データコピー監視機能は図1の場合と一部処理が異なる。記憶サブシステム1が、同期転送により記憶サブシステム2へホストから受領した更新データを転送した後、記憶サブシステム1から3に対し、上述した「データコピー監視機能」を起動させる。つまり、「データ進捗問い合わせ」コマンドを発行し、記憶サブシステム1の持つ転送状態/ビットマップ#1と、記憶サブシステム3の持つ転送状態/ビットマップ#3で、それぞれのUpdateアラフ、カウンタ値を登録し、所定の操作を行う。

【0142】記憶サブシステム1から3に、ホストから記憶サブシステム1が受領したデータ（トラック）と同じデータが、記憶サブシステム3に届いたか否か、問合せた結果、未着であれば、記憶サブシステム1の転送状態/ビットマップ#1のビットマップは、そのまま保持する。結果が到着であれば、つまり、#3のビットマップのブロックナンバ、カウンタ値が同一であれば、Updateアラフを削除し、#1のビットマップを削除する。

【0143】＜同期間におけるその他の処理＞データコピー監視機能で検出した「データ進捗問い合わせ」コマンドの応答結果に、エラーや不具合（タイムアウト）が生じたり、転送状態/ビットマップの機能に不具合が生じた場合には、障害・災害の際に行われるべき回復処理に関する差分管理を禁止する。

【0144】転送状態/ビットマップの機能において、ビットマップは有限なカウンタ値の格納領域を有している。この有限値を超えて（オーバーフロー）、同一データブロックが更新された場合には、そのデータブロックは、その後2以上の記憶サブシステム間で冗長度が維持されていても、災害・障害発生後に再同期処理、差分コピー処理が行なわれる際に、必ず更新対象として扱う。

【0145】正常な運用において直接、データ転送を行なわない記憶サブシステム間で取り扱われる場合（確認コマンド送出）に対し、所定時間、応答が無い場合は、タイムアウトであるとして再同期処理を禁止する。非同期リモートコピーによるペアの再構築処理や、差分データのみの転送する処理を行わず、禁止する。ペアの相手のデータ更新状態を知ることができないため、そのままペアの再度構築処理を行なわしめることは妥当でないからである。

【0146】＜非同期転送におけるデータの整合性の管理＞例えば、ホストが接続する記憶サブシステム1と記憶サブシステム2とが、記憶サブシステム1から記憶サブシステム2にデータを転写する非同期転送で運用されているとする。この場合、もし、記憶サブシステム1におけるデータの書き込み順と、記憶サブシステム2におけるデータの書き込み順とが異なると、同期転送したデータ1、2におけるデータの整合性が保証されなくなる。以下、このようなデータの整合性を回復するための、

仕組みについて説明する。

【0147】まず、各記憶サブシステム1、2における記憶資源の記憶領域に所定サイズ（例えば、16Kバイトごと）のブロックを画し、各ブロックに固有のブロック番号を割り当てる。そして、ホストからデータの書き込みがあったブロックについて、そのブロック番号とデータの書き込み順に付与したシーケンス番号との対応づけを制御メモリ6に管理する。例えば、図12に示すように、ブロック番号が56～59のブロックにデータが書き込まれた場合には、図13に示すデータ管理情報を制御メモリ6に作成する。

【0148】記憶サブシステム1から記憶サブシステム2への非同期転送に際しては、図14の転送データフォーマットに示すように、転送するデータに前記データ管理情報を付帯させる。一方、これを受領した記憶サブシステム2では、図15に示すように、データに付帯して送られてきた前記データ管理情報を、制御メモリ6に管理する。ここで制御メモリ6に管理される前記データの管理情報、すなわち、シーケンス番号とブロックIDの組み合わせは、これに対応するデータのキャッシュメモリ上の位置情報も対応づけ記憶されている。記憶サブシステム2は、前記データ管理情報のシーケンス番号の順番にこれに対応するキャッシュメモリ上の前記位置情報に記憶されているデータを記憶資源に書き込んでいく。

【0149】以上のようにして、ホストが記憶サブシステム1の記憶資源に書き込みた順番とおり、記憶サブシステム2の記憶資源においてもデータが書き込まれ、両記憶サブシステム1、2におけるデータの整合性が保証されることとなる。

【0150】＜マルチホップ方式＞図16（a）に示す広域データストレージシステムは、サイト1に設置された記憶サブシステム1と、サイト2に設置された記憶サブシステム2と、サイト3に設置された記憶サブシステム3とを備える。記憶サブシステム1には、この記憶サブシステム1を記憶手段として利用するホストが接続する。記憶サブシステム1と記憶サブシステム3との間も通信手段により接続される。

【0151】記憶サブシステム1と記憶サブシステム2とは、記憶サブシステム1から記憶サブシステム2にデータを転写する同期転送で運用されている。また、記憶サブシステム2と記憶サブシステム3とは、記憶サブシステム2から記憶サブシステム3にデータを転写する非同期転送で運用されている。以下、このような形態のリモートコピー制御方法を「マルチホップ方式」と称する。なお、マルチホップ方式における各記憶サブシステム間の通信を同期転送とするか、非同期転送とするかは任意に設定される。また、これら以外の転送方式であってもよい。

【0152】つぎに、図16（b）とともにマルチホッ

プ方式によるデータ差分管理の詳細について説明する。
【0153】記憶サブシステム1は、ホストから書き込み対象データとその書き込み要求（Write I/O）を受信すると（S121）、書き込み対象データを自身の論理ボリューム（第1の記憶資源）に書き込むとともに、書き込み処理を行った順にシーケンス番号を付与し、これと前記データが書き込まれた論理ボリューム（第1の記憶資源）上の位置（格納位置）を特定する書き込み位置情報とを対応づけて（所定のテーブルに）記憶する（S122）。なお、書き込み位置情報は、例えば、セクタ番号、トラック番号等を用いて記述される。

【0154】つぎに、記憶サブシステム1は、前記書き込み対象データを、これに付与された前記シーケンス番号とともに記憶サブシステム2に送信する（S123）。ここでこのように記憶サブシステム間で行われる、データとシーケンス番号の送信は、例えば、データ送信コマンドを送信した後に行われ、また、このコマンドには必要に応じて、前述したデータの書き込み位置情報が付帯される。

【0155】記憶サブシステム2は、記憶サブシステム1から送られてくる前記書き込み対象データとシーケンス番号とを受領して、これを自身の論理ボリューム（第2の記憶資源）に書き込み、記憶サブシステム2は、前記書き込み処理が完了すると、その完了通知を記憶サブシステム1に送信する。

【0156】記憶サブシステム2は、記憶サブシステム3に対し、適宜タイミングで前記書き込み対象データと前記シーケンス番号とを送信する（S124）（なお、図16（b）では、時間差を表現するため、記憶サブシステム1から記憶サブシステム2に送信されるデータのシーケンス番号と、記憶サブシステム2から記憶サブシステム3に送信されるデータのシーケンス番号とを交差させている）。

【0157】つぎに、記憶サブシステム3は、前記データと前記シーケンス番号とを受信すると、前記書き込み対象データに对应して発行した前記シーケンス番号を、記憶サブシステム1に送信する（S125）。記憶サブシステム1は、記憶サブシステム3から送られてくるシーケンス番号を受信する。

【0158】ここで記憶サブシステム1は、受信したシーケンス番号と、自分が記憶しているシーケンス番号とこれに对应する書き込み位置情報との対応づけ（テーブル）を対照することで、記憶サブシステム3の論理ボリューム（第3の記憶資源）に未反映のデータ、すなわち、差分データを把握することができる。なお、前記の対照は、例えば、記憶サブシステム3から受領した書き込み完了位置までのシーケンス番号と書き込み位置情報とをテーブルから削除することにより行われる（S126）。

【0159】以上のようにしてマルチホップ方式にお

る通常運用が行われる。

【0160】 つぎに、災害等により記憶サブシステム2が停止した場合の回復処理について説明する。

【0161】 図17 (a) に示すように、記憶サブシステム1は、例えば、ハートビートメッセージの監視などの障害検出機能により、記憶サブシステム2の稼働状態をリアルタイムに監視している。以下では、ハートビートメッセージが検出されるなどして、記憶サブシステム1が記憶サブシステム2の障害発生を検知した場合に、記憶サブシステム1と記憶サブシステム3の間を、差分データののみを複写することによってその内容を一致させ、その後記憶サブシステム1と記憶サブシステム3の間を、非同同期転送での臨時運用へ移行させる処理について、図17 (b) とともに説明する。

【0162】 記憶サブシステム1は、記憶サブシステム2の障害発生を検知した場合 (S131)、まず、制御メモリ6上に、自身の論理ボリューム (第1の記憶資源) の所定ブロック単位のデータ格納位置に対応づけたビットマップを生成し、自身が記憶している記憶サブシステム3において未反映の前記差分データについての前記シーケンス番号と前記書き込み位置情報とを対応づけて、データ更新のあった前記ビットマップに対応する位置のビットをオンにする (S132)。

【0163】 つぎに、記憶サブシステム1の論理ボリュームの、前記ビットマップ上のオンになっている位置に格納されている差分データを記憶サブシステム1から記憶サブシステム3の対応する格納位置に複写する (S133)。そして、この複写完了後、記憶サブシステム1から非同同期転送により差分データが複写される状態で、臨時運用が開始される (S134)。

【0164】 ここでこの臨時運用への切り替えに際しては、記憶サブシステム2に障害が発生した場合でも、記憶サブシステム1のデータを記憶サブシステム3に全部複写する必要がなく、差分データのみを複写すればよい。このため、例えば、記憶サブシステム1と記憶サブシステム3との間の通信回線のデータ伝送量が充分でない場合でも、各記憶サブシステムにおける論理ボリュームに記憶されているデータを容易に同期させることができる。

【0165】 つぎに、記憶サブシステム2が復旧し、臨時運用から通常運用に切り替える際の処理について説明する。

【0166】 まず、記憶サブシステム1は、自身の論理ボリューム (第1の記憶資源) に記憶されている全てのデータを記憶サブシステム2の論理ボリューム (第2の記憶資源) に複写した後、記憶サブシステム1から記憶サブシステム2にデータを複写する同期転送での運用を開始する。すなわち、記憶サブシステム1は、ホストからの指示により自身の論理ボリューム (第1の記憶資源) にデータ書き込みを行った場合、書き込んだデータとシ

ーケンス番号とを記憶サブシステム2に送信する。

【0167】 記憶サブシステム2は、記憶サブシステム1から送られてくる前記書き込んだデータとシーケンス番号とを受信して、これを自身の論理ボリューム (第2の記憶資源) に書き込む。記憶サブシステム2は、前記書き込み処理が完了すると、自身の論理ボリューム (第2の記憶資源) に対するデータ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシーケンス番号とを対応づけて (所定のテーブルに) 記憶する。この段階のデータ転送状態を図18に示す。

【0168】 つぎに、記憶サブシステム3は、記憶サブシステム1から送られてくる前記データと前記シーケンス番号とを受信して、前記データを自身の論理ボリューム (第3の記憶資源) に記憶するとともに (図18) 前記対応づけにおける前記シーケンス番号を記憶サブシステム2に送信する (図示せず)。

【0169】 記憶サブシステム2は、記憶サブシステム3から送られてくるシーケンス番号を受信する。ここで記憶サブシステム2は、前記受信したシーケンス番号と、自身が記憶しているシーケンス番号と、これに対応する書き込み位置情報とを対照することで、記憶サブシステム3の論理ボリュームに未反映のデータ、すなわち、差分データを把握することができる。

【0170】 つぎに、臨時運用において記憶サブシステム1から記憶サブシステム3に複写する非同同期転送の運用を停止する。この停止後、記憶サブシステム2は、自身の制御メモリ上に、自身の論理ボリューム (第2の記憶資源) の所定ブロック単位のデータ格納位置に対応づけたビットマップを生成し、自身が記憶している記憶サブシステム3において未反映の前記差分データについてシーケンス番号と書き込み位置情報との対応づけに基づいて、データ更新のあった前記ビットマップの該当位置のビットをオンにする。

【0171】 つぎに、記憶サブシステム2は、前記ビットマップにより把握した、記憶サブシステム3の論理ボリューム (第3の記憶資源) において未反映となっている差分のデータとその書き込み位置情報とを記憶サブシステム3に送信する。

【0172】 記憶サブシステム3は、前記差分データと前記書き込み位置情報とを受信して、前記差分データを、自身の論理ボリューム (第3の記憶資源) の、前記書き込み位置情報により指定される該当データの格納位置に記憶する。これにより、記憶サブシステム2の論理ボリューム (第2の記憶資源) の内容と、記憶サブシステム3の論理ボリューム (第3の記憶資源) の内容との同期が取れることになる。以上の処理終了後、記憶サブシステム2と記憶サブシステム3との間の非同同期転送による運用が開始され、図19に示す通常状態での運用が再開する。

【0173】 以上のようにして臨時運用から通常運用への切り替えが完了する。

【0174】 <マルチコピー方式>図20に示す広域データストレージシステムは、サイト1に設置された記憶サブシステム1と、サイト2に設置された記憶サブシステム2と、サイト3に設置された記憶サブシステム3とを備える。記憶サブシステム2にはこの記憶サブシステム2を記憶手段として利用するホストが接続する。なお、記憶サブシステム1と記憶サブシステム3との間も通信手段により接続される。

【0175】 記憶サブシステム1と記憶サブシステム2とは、記憶サブシステム2から記憶サブシステム1にデータを複写する同期転送で運用されている。また、記憶サブシステム2と記憶サブシステム3とは、記憶サブシステム2から記憶サブシステム3にデータを複写する非同同期転送で運用されている。以下、このような形態のリモートコピー制御方法を「マルチコピー方式」と称する。なお、マルチコピー方式において、各記憶サブシステム間の通信を同期転送とするか、非同同期転送とするかは前記の形態に限らず、任意に設定される。また、同期転送や非同同期転送以外の転送方式であってもよい。

【0176】 つぎに、図20とともにこの実施例のデータ差分管理方式について説明する。記憶サブシステム2は、ホストから書き込み対象データとその書き込み要求 (Write 1/0) とを受信すると (S161)、書き込み対象データを自身の論理ボリューム (第2の記憶資源) に書き込む。また、記憶サブシステム2は、書き込まれたデータと、書き込み処理を行った順に付与したシーケンス番号とを、記憶サブシステム1に送信する (S162)。そして同時に、前記書き込まれたデータと前記付与したシーケンス番号とを、記憶サブシステム3に送信する (S164)。なお、前述のマルチホップ方式の場合と同様に、このような記憶サブシステム間で行われるデータとシーケンス番号の送信は、例えば、データ送信コマンドを送信した後に行われ、また、このコマンドには、必要に応じて前述したデータの書き込み位置情報が付帯される。

【0177】 つぎに、記憶サブシステム1は、記憶サブシステム2から送られてくる前記書き込み対象データとシーケンス番号とを受信して、前記書き込み対象データを自身の論理ボリューム (第1の記憶資源) に書き込む。その際、前記シーケンス番号と、これと前記データが書き込まれた論理ボリューム (第1の記憶資源) 上の位置 (格納位置) を特定する書き込み位置情報とを対応づけて (所定のテーブルに) 記憶する (S163)。なお、書き込み位置情報は、例えば、セクタ番号、トラック番号等を用いて記述される。

【0178】 つぎに、記憶サブシステム3は、記憶サブシステム2から送られてくる前記書き込み対象データとシーケンス番号とを受信して、前記書き込み対象データを自身の論理ボリューム (第3の記憶資源) に書き込

む。書き込みが完了すると、記憶サブシステム3は記憶サブシステム1に対し、前記書き込み対象データとこれと対応していた前記シーケンス番号とを記憶サブシステム1に送信する (S165)。記憶サブシステム1は、記憶サブシステム3から送られてくるシーケンス番号を受信する。

【0179】 ここで記憶サブシステム1は、前記受信したシーケンス番号と、自身が記憶しているシーケンス番号とこれと対応する書き込み位置情報との対応づけを対照することで、記憶サブシステム3の論理ボリューム (第3の記憶資源) に未反映のデータ、すなわち、差分データを把握することができる。なお、前記の対照は、例えば、記憶サブシステム3から受領した書き込み完了位置までのシーケンス番号と書き込み位置情報とをテーブルから削除することで行われる (S166)。

【0180】 以上のようにしてマルチコピー方式における通常運用が行われる。

【0181】 つぎに、災害等により記憶サブシステム2が停止した場合の回復処理について説明する。

【0182】 図21 (a) に示すように、記憶サブシステム1は、例えば、ハートビートメッセージの監視などの障害検出機能により、記憶サブシステム2の稼働状態をリアルタイムに監視している。以下では、ハートビートメッセージが検出されるなどして、記憶サブシステム1が記憶サブシステム2の障害発生を検知した場合に、記憶サブシステム2に接続するホストに代えて、記憶サブシステム1と記憶サブシステム3の間を、差分データののみを複写することによってその内容を一致させ、その後記憶サブシステム1と記憶サブシステム3の間を、非同同期転送での臨時運用へ移行させる処理について、図21 (b) とともに説明する。

【0183】 記憶サブシステム1は記憶サブシステム2の障害発生を検知した場合 (S171)、例えば、オペレータの操作により、記憶サブシステム2に接続していたホストの業務の運用が、記憶サブシステム1に接続する副ホストに引き継がれる。

【0184】 つぎに、記憶サブシステム1は、制御メモリ6上に、自身の論理ボリューム (第1の記憶資源) の所定ブロック単位のデータ格納位置に対応づけたビットマップを生成し、自身が記憶している記憶サブシステム3において未反映の前記差分データについてシーケンス番号とデータ更新位置情報との対応づけに基づいて、データ更新のあった前記ビットマップの該当位置のビットをオンにする (S172)。

【0185】 つぎに、記憶サブシステム1の論理ボリュームの、前記ビットマップ上のオンになっている位置に対応する位置に格納されている差分データを、記憶サブシステム1から記憶サブシステム3に複写する (S173)。そして、複写完了後、記憶サブシステム1から同期転送によりデータが複写される状態で、臨時運用が開

始される (S174)。

【0186】 ここでこの臨時運用への切り替えに際しては、記憶サブシステム2に障害が発生した場合でも、記憶サブシステム1のデータを記憶サブシステム3に全部複写する必要がなく、差分データのみを複写すればよい。このため、例えば記憶サブシステム1と記憶サブシステム3との間の通信回線のデータ伝送量が充分でない場合でも、各記憶サブシステムにおける論理ボリュームに記憶されているデータを簡単に同期させることができる。

【0187】 つぎに、記憶サブシステム2が復旧し、臨時運用から通常運用に切り替える際の一連の処理について説明する。

【0188】 まず、記憶サブシステム1は、自身の論理ボリューム (第1の記憶資源) に記憶している全てのデータを記憶サブシステム2の論理ボリューム (第2の記憶資源) に複写した後、記憶サブシステム1から記憶サブシステム2にデータを複写する同期転送での運用を開始する。なお、このとき記憶サブシステム1と記憶サブシステム3間での非同期転送も継続して行われる。

【0189】 記憶サブシステム1は、ホストから書き込まれたデータと、書き込み処理を行った後に付与したシークエンス番号とを、記憶サブシステム2に送信する。そして同時に、前記書き込まれたデータと前記付与したシークエンス番号とを、記憶サブシステム3にも送信する。

【0190】 記憶サブシステム2は、自身の論理ボリューム (第2の記憶資源) に対するデータ書き込みが行われた位置を特定する書き込み位置情報と、データの書き込み順に付与されるシークエンス番号との対応づけを記憶する (位置情報管理テーブル作成)。この段階での運用状態を図22に示す。

【0191】 記憶サブシステム3は、記憶サブシステム1から送られてくる前記データと前記シークエンス番号とを受信して、前記データを自身の論理ボリューム (第3の記憶資源) に記憶するとともに前記対応づけにおける前記シークエンス番号を記憶サブシステム2に送信する。

【0192】 記憶サブシステム2は、記憶サブシステム3から送られてくるシークエンス番号を受信する。ここで記憶サブシステム2は、前記受信したシークエンス番号と、自身が記憶している前記対応づけを対照すること、自身が記憶している前記対応づけを対照すること、で、記憶サブシステム3の論理ボリュームに未反映のデータ、すなわち、差分データを把握することができる。

【0193】 つぎに、臨時運用において記憶サブシステム1から記憶サブシステム3に複写する非同期転送の運用を停止する。この停止後、記憶サブシステム2は、自身の論理メモリ上に、自身の論理ボリューム (第2の記憶資源) の所定ブロック単位のデータ格納位置に対処したビットマップを生成し、自身が記憶している記憶サブシステム3において未反映の前記差分データについてシークエンス番号と書き込み位置情報との対応づけに基

づいて、データ更新のあった前記ビットマップの該当位置のビットをオンにする。

【0194】 つぎに、記憶サブシステム2は、前記ビットマップにより把握した、記憶サブシステム3の論理ボリューム (第3の記憶資源) において未反映となっている差分のデータとその書き込み位置情報とを記憶サブシステム3に送信する。

【0195】 記憶サブシステム3は、前記差分データと前記書き込み位置情報とを受信して、前記差分データを前記書き込み位置情報に基づいて自身の論理ボリューム (第3の記憶資源) に記憶する。これにより、記憶サブシステム2の論理ボリューム (第2の記憶資源) の内容と、記憶サブシステム3の論理ボリューム (第3の記憶資源) の内容との同期が取れることになる。それから記憶サブシステム2から記憶サブシステム3への非同期転送が開始される。この段階での運用状態を図23に示す。

【0196】 ここで記憶サブシステム1に接続するホストの記憶サブシステム1へのデータ書き込み処理が完了しており、記憶サブシステム1と記憶サブシステム2の同期が取れている時に、記憶サブシステム1から記憶サブシステム2に対して行っていたデータの複写を、記憶サブシステム2から記憶サブシステム1に対して行うように切り替える。すなわち、同期が取れている状態で切り替えを行うことで、差分データを複写する等の作業が必要でなくなる。

【0197】 つぎに、記憶サブシステム1に接続するホストにより運用されている業務を、記憶サブシステム2に接続するホストに引き継ぐ。そして、記憶サブシステム2から記憶サブシステム3にデータを複写する同期転送による運用を開始することで、図24に示す通常状態での運用が再開することになる。

【0198】 以上のようにして臨時運用から通常運用への切り替えが完了する。

【0199】 <他の障害復旧方式> つぎに、障害復旧方式のバリエーションについて説明する。

【0200】 図25に示すマルチホップ方式において、記憶サブシステム1がダウンした場合 (a) には、記憶サブシステム2に副ホストを接続し、この副ホストにより記憶サブシステム1に接続するホストの業務を引き継ぐ。なお、記憶サブシステム2と記憶サブシステム3の間では、非同期転送での運用が行われている (b)。

【0201】 記憶サブシステム1が復旧した場合には、まず、記憶サブシステム2の全データを記憶サブシステム1に複写し、副ホストの業務を記憶サブシステム1に接続するホストに引き継ぐ。そして、前記の要領で、記憶サブシステム1と記憶サブシステム2との間のデータ転送方向を逆向きにすることにより、通常運用を再開する (c)。

【0202】 図26に示すマルチホップ方式において、

記憶サブシステム3に障害が発生した場合 (a) には、記憶サブシステム3の復旧後、記憶サブシステム2から記憶サブシステム3に全データを複写して記憶サブシステム3のデータを記憶サブシステム2と同期させ、記憶サブシステム1から記憶サブシステム2にデータを複写する同期転送および記憶サブシステム2から記憶サブシステム3にデータを複写する非同期転送による通常運用を再開する (b)。

【0203】 図27に示すマルチホップ方式において、記憶サブシステム1に障害が発生した場合 (a) には、記憶サブシステム1の復旧後、記憶サブシステム2から記憶サブシステム1に全データを複写して記憶サブシステム1のデータを記憶サブシステム2と同期させ、記憶サブシステム2から記憶サブシステム1にデータを複写する同期転送および記憶サブシステム2から記憶サブシステム3にデータを複写する非同期転送による通常運用を再開する (b)。

【0204】 図28に示すマルチホップ方式において、記憶サブシステム3に障害が発生した場合には、記憶サブシステム3の復旧後、記憶サブシステム2から記憶サブシステム3に全データを複写して記憶サブシステム3のデータを記憶サブシステム2と同期させ、記憶サブシステム2から記憶サブシステム1にデータを複写する同期転送および記憶サブシステム2から記憶サブシステム3にデータを複写する非同期転送による通常運用を再開する (b)。

【0205】 <複写元・複写先、書き込み位置情報の管理> 記憶サブシステム間でデータを転送する場合、データの転送元や転送先の設定や、その転送が同期・非同期いずれかの方式で行われるといった設定は、オペレータが各記憶サブシステムを操作して設定する場合 (なお、この場合には、例えば、ある記憶サブシステムが障害を起して使えなくなった場合に、どの記憶サブシステムが次のデータの転送元になり、どの記憶サブシステムが次の転送先になるのかということも、システム構成時に予め登録しておく)、記憶サブシステムに付随するシステムが自動的に行うようにしている場合など、システムの構成に応じて様々な形態で行われる。

【0206】 また、シークエンス番号と書き込み位置情報との対応づけの管理は、例えば、オペレータが、転送元や転送先を記憶サブシステムに登録する操作を開始する機器で行う。

【0207】 <記憶サブシステムの選択方式> 図29に示す広域データストレージシステムは、記憶サブシステム1とこれに接続するホスト1h、記憶サブシステム1からデータが非同期転送される記憶サブシステム2および記憶サブシステム3を備えている。ホスト1hもしくは記憶サブシステム1に障害が発生した場合、迅速に記憶サブシステム2もしくは記憶サブシステム3のどちらか一方を主たる記憶サブシステムとして選択し、また、

信頼性・良全性確保のため、これら2つの記憶サブシステム2および3においてデータを2重化管理する。以下、ホスト1hもしくは記憶サブシステム1に障害が発生した場合に行われる処理について説明する。

【0208】 記憶サブシステム2は、例えば、記憶サブシステム1から送信されてくるデータの有無や、記憶サブシステム1からあらかじめ設定された時間等に送られてくるハートビートメッセージの監視により、ホスト1hや記憶サブシステムに障害が発生したことを検知する。

【0209】 障害を検知した場合、記憶サブシステム2は、迅速に主たる記憶サブシステムを決定し、副ホストもしくは副ホスト3による臨時運用に切り替える。主たる記憶サブシステムの選択はつぎのようにして行われる。まず、障害を検知した記憶サブシステム2は、記憶サブシステム3に、前述したシークエンス番号のうち最新のシークエンス番号の送信を要求するメッセージを送信する。記憶サブシステム3は、前記メッセージを受信すると、自身が記憶している最新のシークエンス番号を記憶サブシステム2に送信する。

【0210】 記憶サブシステム2は、記憶サブシステム3から送られてきたシークエンス番号と、自身が記憶している最新のシークエンス番号とを比較して、より最新のシークエンス番号を受信している記憶サブシステムを、主たる記憶サブシステムとして選出し、選出した記憶サブシステムの識別子を選出候補として記憶するとともに、前記識別子を記憶サブシステム3に送信する。記憶サブシステム3は、送信されてきた前記識別子を受信し、これによりどの記憶サブシステムが主たるサブシステムとして選出されたのかを認知する。

【0211】 なお、以上の選出処理において、記憶サブシステム間の通信方式の性質などの諸事情により、記憶サブシステム2もしくは記憶サブシステム3が記憶しているシークエンス番号に抜けが存在しているシークエンス番号で、このような場合には、継続しているシークエンス番号のうちで最新のものを、前記の比較に用いる。

【0212】 主たる記憶サブシステムが選出されると、つぎに、記憶サブシステム2と記憶サブシステム3とによりデータの二重化管理を行うため、両者が記憶しているデータの内容を一致させる。これは、記憶サブシステム間で全データの複写や差分データの複写により行われる。記憶サブシステム間でデータが一致すると、主たる記憶サブシステムとして選出された記憶サブシステムは、自身に接続している副ホストに、自身が主たる記憶サブシステムとなる旨を送信する。副ホストはこれを受信して代行運用を開始する。また、記憶サブシステム2と記憶サブシステム3との間で、同期転送もしくは非同期転送によるデータの二重化管理が開始される。

【0213】 なお、以上の説明では、記憶サブシステム2が記憶サブシステム3から最新のシークエンス番号を取

得て、主たる記憶サブシステムを選出するようにしているが、この処理は記憶サブシステム3が行ってもよい。

【0214】また、記憶サブシステム1乃至記憶サブシステム3の3台構成の記憶サブシステムにおいて、記憶サブシステム1の障害発生時に代行して運用される他の記憶サブシステムを選出する仕組みを一例として説明したが、前述の仕組みは、4台以上の記憶サブシステムで構成される広域データストレージシステムにも適用することができ。

【0215】＜キャッシュメモリ上のデータの管理＞ホストが接続する一次の記憶サブシステムに、この一次の記憶サブシステムのデータのリモートコピ先である1以上の二次の記憶サブシステムが接続する系における、一次の記憶サブシステムのキャッシュメモリ上のデータの管理に関する実施例について説明する。

【0216】前記の系において、一次の記憶サブシステムから二次の記憶サブシステムに複写（リモートコピ）する必要の無いデータについては、一次の記憶サブシステムの記憶資源にデータを書き込んだ後は、そのデータを当該記憶サブシステムのキャッシュメモリに複写する。また、二次の記憶サブシステムに複写する場合には、少なくともそのデータを二次の記憶サブシステムに送付するまではキャッシュメモリ上に残しておく必要がある。また、転送先となる二次の記憶サブシステムが複写存在する場合には、通信手段の違いや運用上の差異などにより、通常、二次の記憶サブシステムについての転送が同時に行われるわけではないので、このような場合には、全ての二次の記憶サブシステムに対する転送が終了するまで、データをキャッシュメモリ上に残しておく仕組みが必要である。

【0217】そこで、一次の記憶サブシステムにおいて、キャッシュメモリ上に置かれているデータについて、一次の記憶サブシステムに接続する二次の各記憶サブシステムについての転送が完了しているかどうかを管理するようにする。具体的には、例えば、図30に示すように、キャッシュメモリ上に区画された記憶ブロック（＃1、～、＃n）ごとに、それぞれの記憶ブロックに格納されているデータについて、二次の各記憶サブシステムへの転送が完了しているかどうかを示すテーブルを、一次の記憶サブシステムにおいて管理するようにする。

【0218】なお、このテーブルにおいて、ビット「0」は転送が完了していることを示し、ビット「1」は転送が完了していないことを示す。ホストからのデータが一次の記憶サブシステムに書き込まれた時に、データが書き込まれた記憶ブロックの転送先となっている二次の記憶サブシステムに対応するビットに「1」がセットされる。ある記憶ブロックの「1」がセットされているビットのうち、データの転送が完了した二次の記憶サブシステムについてのビットは、転送完了後に「0」と

なる。

【0219】そして、全ての二次の記憶サブシステムについてのビットが「0」となった記憶ブロックに格納されているデータについては、キャッシュメモリ上から消去してもよいということになる。

【0220】

【発明の効果】図1、図9及び図10で示した、3つ以上のサイトを有する広域データストレージシステムにおいて、いずれかのサイトに、いつ災害・障害が発生しても、巨额的に見て常時、データの順序性を保証した論理ボリュームを残すことができる。

【0221】直接データ転送に図与しなかった論理ボリューム間、例えば、図7の記憶サブシステム1と記憶サブシステム3との間で、差分データのみをコピーすれば、即時に、非同期リモートコピのメプを生成でき、広域データストレージシステムの運用再始が、即時に、可能となる効果がある。

【0222】本発明では、記憶サブシステムの内部にリモートコピを実施するための互換性論理ボリュームを必要としなため、記憶サブシステムのメモリ資源の使用効率が上がり、記憶サブシステムのコストパフォーマンスが向上する効果がある。

【図面の簡単な説明】

【図1】本発明に係る広域データストレージシステムの全体構成の一例を示した説明図である。

【図2】記憶サブシステムの一例を示した概念図である。

【図3】図1の構成において、データコピ監視機能を説明するための概念図である。

【図4】本発明を実現するための転送状態/ビットマップの一例を示した図である。

【図5】一般的な非同期転送によるコピの制御の概略を説明するための図である。

【図6】非同期リモートコピの制御の概略を説明するための図である。

【図7】図9の全体構成において、データセンタ2に障害・災害が発生した場合の復旧の様子を示した説明図である。

【図8】図1の全体構成において、データセンタ1に障害・災害が発生した場合の復旧の様子を示した説明図である。

【図9】本発明に係る広域データストレージシステムの全体構成の別の一例を示した説明図である。

【図10】データセンタを4拠点以上設置した場合、本発明に係る広域データストレージシステムの全体構成の別の一例を示した説明図である。

【図11】図9の全体構成において、データコピ監視機能を説明するための概念図である。

【図12】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、記憶資源

のデータを管理する単位であるブロックの概念を示す図である。

【図13】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、データ管理情報の概念を示す図である。

【図14】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、転送データのフォーマットの概念を示す図である。

【図15】本発明の一実施例による非同期転送におけるデータの整合性の管理方法を説明するための、記憶サブシステム2において管理されるデータ管理情報の概念を示す図である。

【図16】（a）はマルチホップ方式の広域データストレージシステムの概念を示す図であり、（b）は（a）に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図17】（a）はマルチホップ方式の広域データストレージシステムの概念を示す図であり、（b）は（a）に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図18】マルチホップ方式において、臨時運用から通常運用に切り替え途中段階における記憶サブシステム間のデータ転送状態を示す図である。

【図19】マルチホップ方式において、臨時運用から通常運用への切り替え終了後の記憶サブシステム間のデータ転送状態を示す図である。

【図20】（a）はマルチホップ方式の広域データストレージシステムの概念を示す図であり、（b）は（a）に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図21】（a）はマルチホップ方式の広域データストレージシステムの概念を示す図であり、（b）は（a）に記載の記憶サブシステムにより行われる処理の流れを示す図である。

【図22】マルチコピ方式において、臨時運用から通常運用に切り替え途中段階における記憶サブシステム間のデータ転送状態を示す図である。

【図23】マルチコピ方式において、臨時運用から通常運用に切り替え途中段階における記憶サブシステム間のデータ転送状態を示す図である。

【図24】マルチコピ方式において、臨時運用から通常運用への切り替え終了後の記憶サブシステム間のデータ転送状態を示す図である。

【図25】（a）～（c）は、マルチホップ方式における障害復旧方式の他のバリエーションを説明する図である。

【図26】（a）、（b）は、マルチホップ方式における障害復旧方式の他のバリエーションを説明する図である。

【図27】（a）、（b）は、マルチコピ方式における障害復旧方式の他のバリエーションを説明する図である。

【図28】（a）、（b）は、マルチコピ方式における障害復旧方式の他のバリエーションを説明する図である。

【図29】障害発生時において、本番業務を代行させる記憶サブシステムの選択方法を説明する、広域データストレージシステムの概念図である。

【図30】本発明の一実施例による、キャッシュメモリ上のデータの管理方法における、二次の各記憶サブシステムへのデータの転送状態を管理するテーブルを示す図である。

【符号の説明】

30 1 記憶サブシステムの制御装置

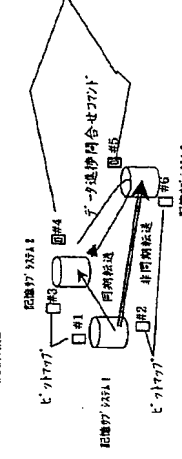
5 キャッシュメモリ

6 制御メモリ

＃1～＃6 転送状態/ビットマップ。

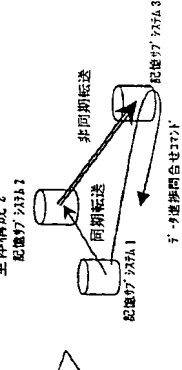
【図3】

データコピ監視機能

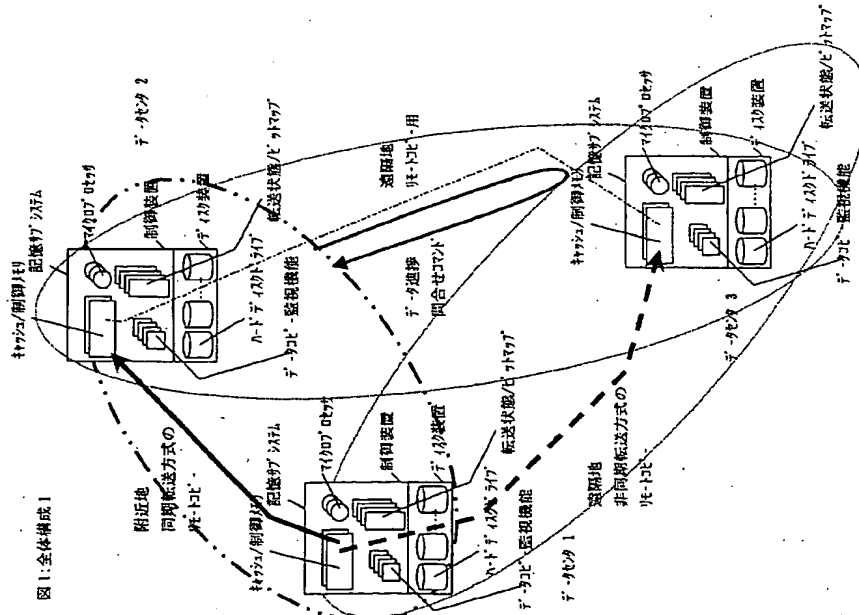


【図9】

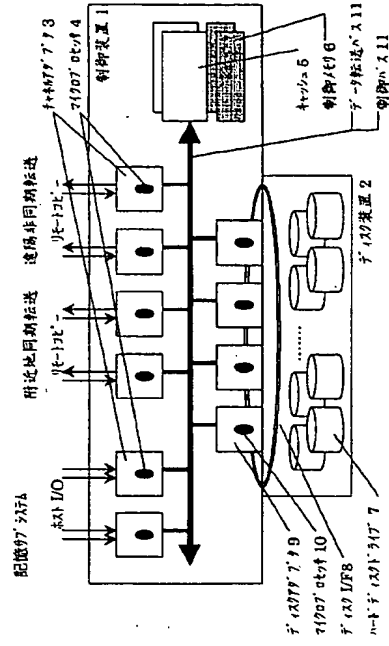
全体構成2



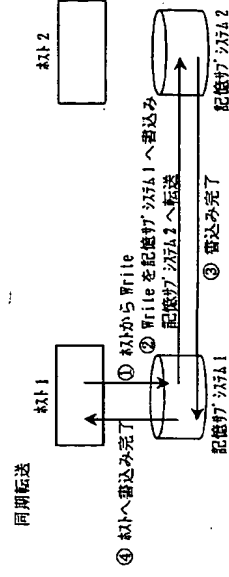
【図1】



【图2】



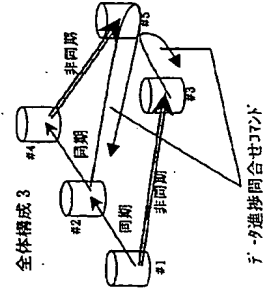
【例5】



【圖 13】

シ・カス番号	0
ﾌﾞﾛｯｸ番号	5 6
シ・カス番号	1
ﾌﾞﾛｯｸ番号	5 7
シ・カス番号	2
ﾌﾞﾛｯｸ番号	5 8
シ・カス番号	3
ﾌﾞﾛｯｸ番号	5 9

【010】



【图 15】

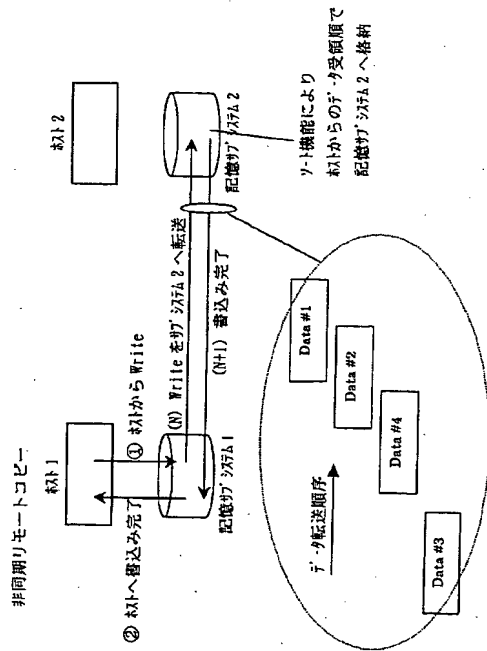
シークス番号	0
ブロッグ番号	56
キヤパシ番号	122
シークス番号	3
ブロッグ番号	59
キヤパシ番号	2
シークス番号	1
ブロッグ番号	57
キヤパシ番号	16

転送状態/ビットマップ

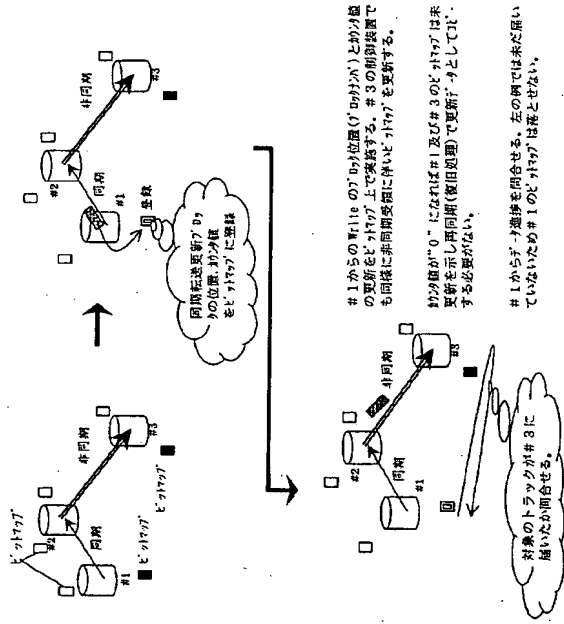
Update #11	Update #12	Update #13	Update #14
3630	1	3	1
3631	0	0	1
3632	1	2	1
4032	1	Over Flow	1

【图4】

【図6】

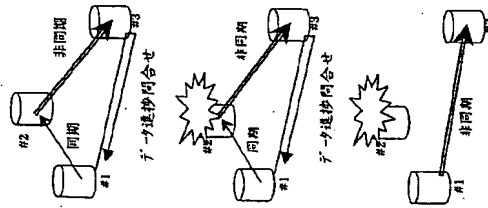


【図11】



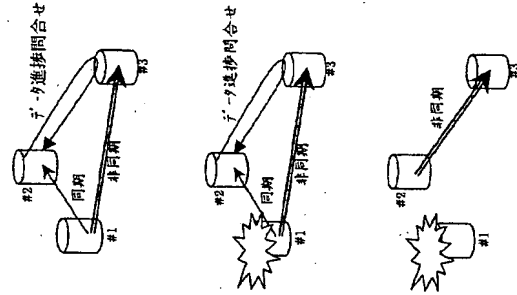
【図7】

障害時の運用と効果



【図8】

障害時の運用と効果



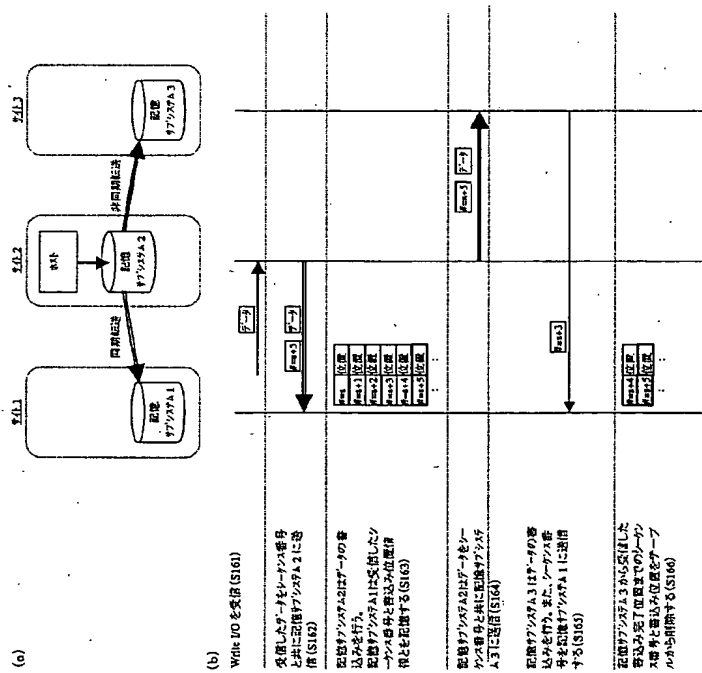
【図12】

プロダクト番号56	プロダクト番号57	プロダクト番号58	プロダクト番号59
-----------	-----------	-----------	-----------

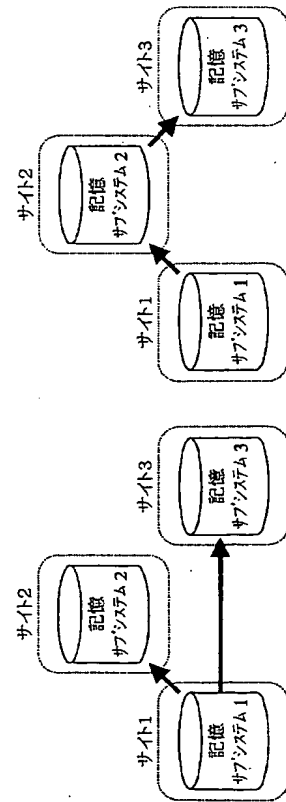
【図14】

プロダクト番号	プロダクト番号	プロダクト番号	プロダクト番号
シリアル番号	シリアル番号	シリアル番号	シリアル番号

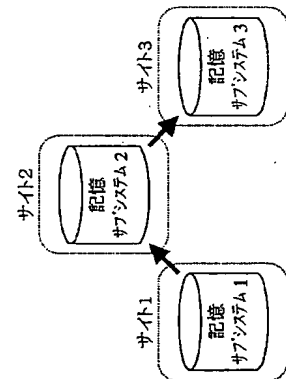
【図 16】



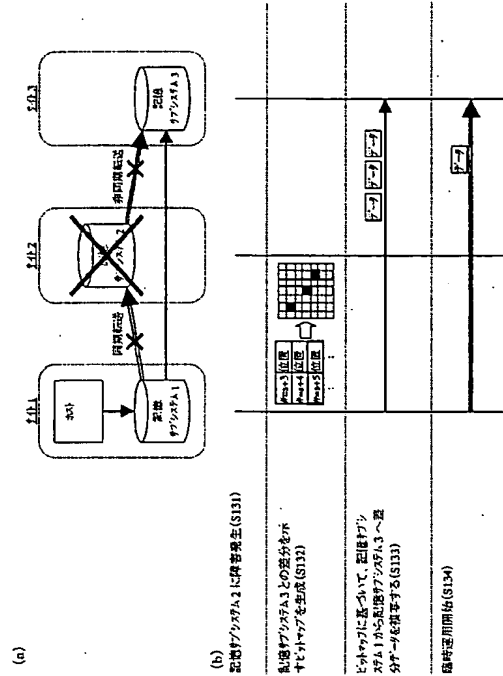
【図 18】



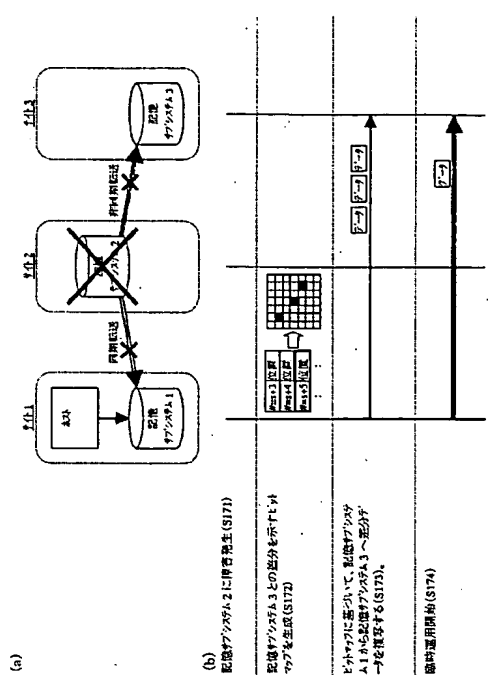
【図 19】



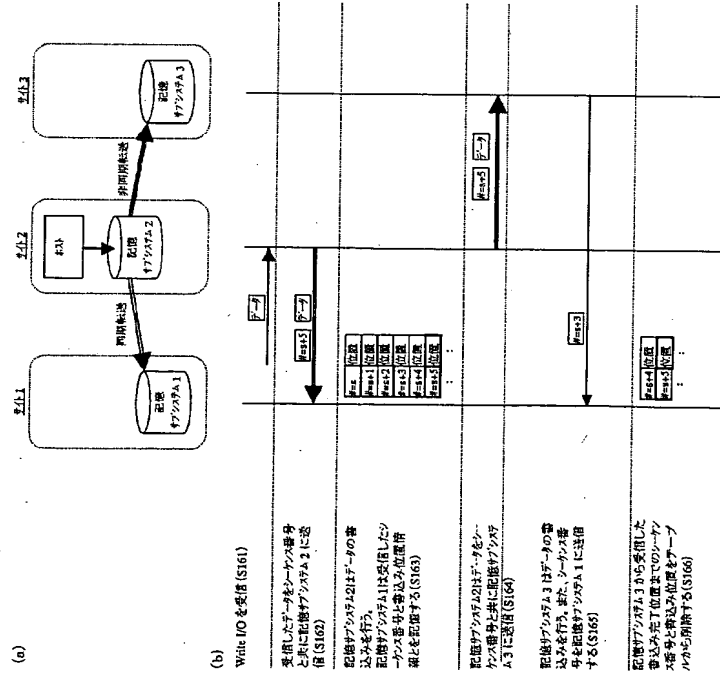
【図 17】



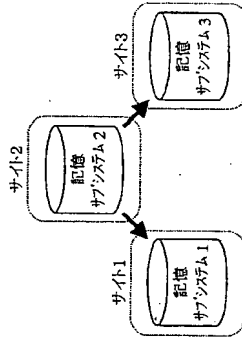
【図 21】



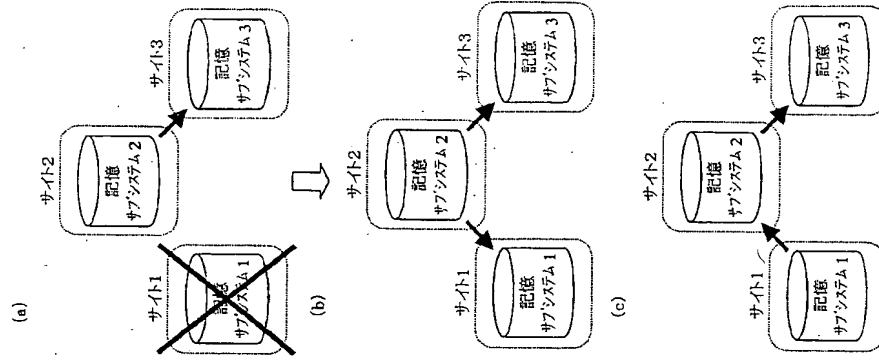
【図20】



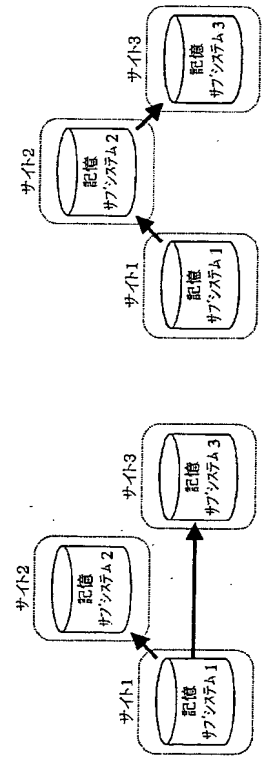
【図24】



【図25】

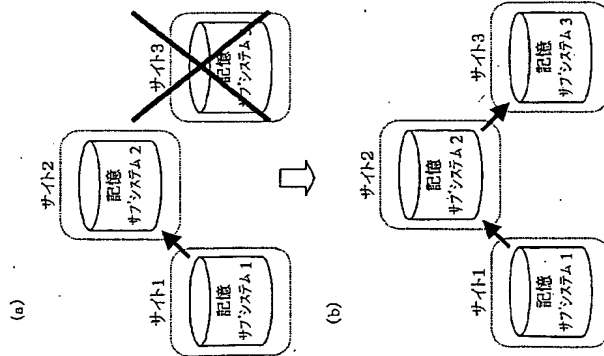


【図22】

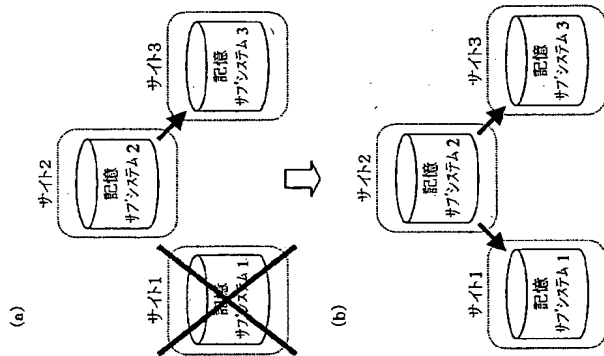


【図23】

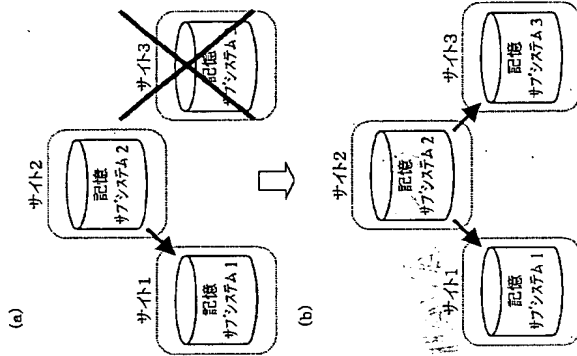
【図 26】



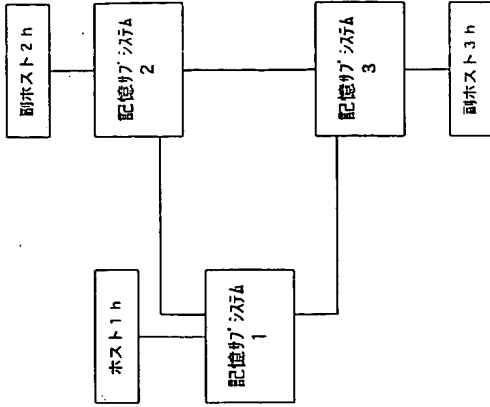
【図 27】



【図 28】



【図 29】



【図 30】

二次データ記憶装置#1	1	0	0	0
二次データ記憶装置#2	1	1	1	1
二次データ記憶装置#3	1	0	0	0
二次データ記憶装置#4	1	1	0	0

フロントページの続き

- (72)発明者 尾形 幹人
神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内
- (72)発明者 岡見 吉規
神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内
- (72)発明者 神垣 誠一
神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内
- (72)発明者 安部井 大
神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内
- (72)発明者 木城 茂
神奈川県小田原市中里322番地2号 株式会社日立製作所RAIDシステム事業部内
- Fターム(参考) 5B005 B401 CE01 EA35
5B082 DE03 GA04

BEST AVAILABLE COPY